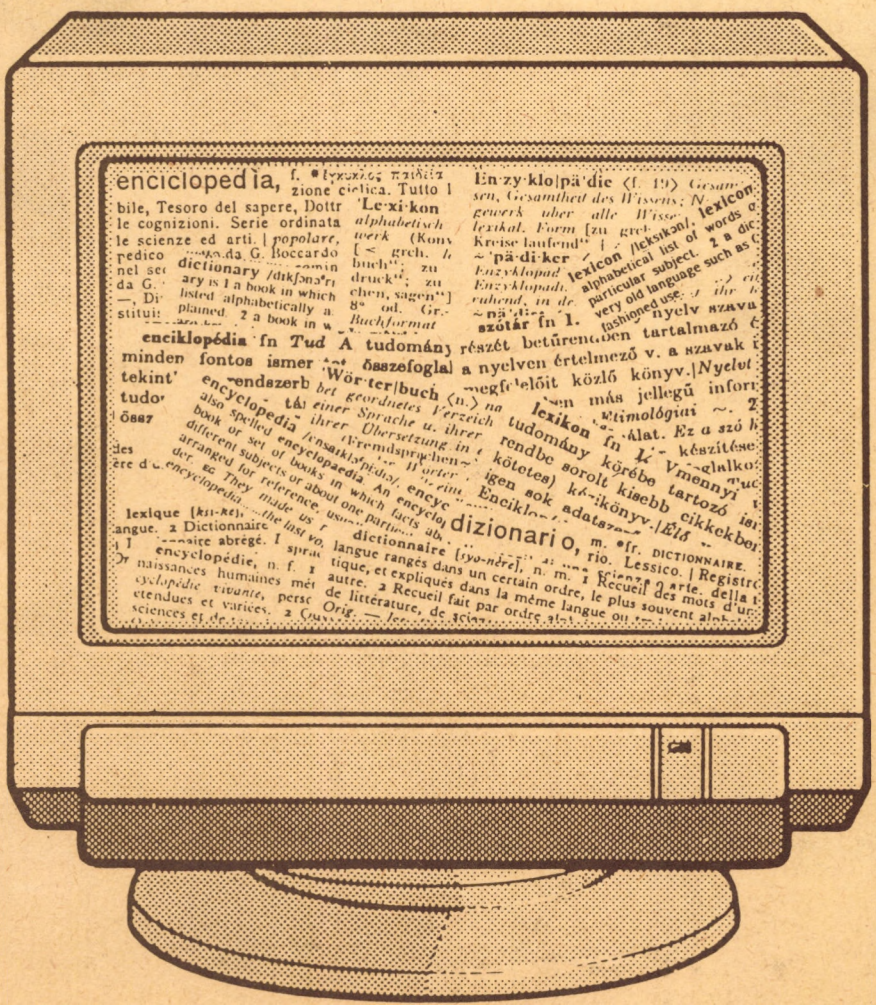
PAPERS
IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX '99

Edited by
Ferenc Kiefer, Gábor Kiss and Júlia Pajzs



LINGUISTICS INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

PAPERS IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX '99

PAPERS
IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX '99

Edited by
Ferenc Kiefer, Gábor Kiss and Júlia Pajzs

Errata

1. Page 154 is the first page of the article of J.Tió—J.M.Cots—M.Sabaté—G.Vázquez—F.Manya—T.Alsinet: "The LVBAC project ..." This page should be inserted between pages 179 and 180.

2. On page 211 the authors of the article "Markup Enhancement ..." should correctly read as Nancy Ide and Tomaž Erjavec.

LINGUISTICS INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

Proceedings of the 4th International Conference on
Computational Lexicography, COMPLEX '99
Budapest, Hungary

All correspondence should be sent to

Linguistics Institute, Hungarian Academy of Sciences
Department of Lexicography and Lexicology
Budapest P.O. Box 19
Hungary 1250

Cover design by Gábor Kiss

ISBN 963 9074 20 9

© Linguistics Institute, Hungarian Academy of Sciences 1999

Contents

Preface

This volume contains the papers presented at the Fifth International Conference on Computational Lexicography and Text Research, organised jointly by the Research Institute for Linguistics of the Hungarian Academy of Sciences and the Laboratoires d'Automatique Documentaire et Linguistique of Université Paris 7. The conference received organisational help from Janus Pannonius University Pécs, and was supported by EURALEX.

The invited speaker Thierry Fontenelle has given an extensive overview of the current trends in corpus analysis and its relation to natural language processing. A set of papers is devoted to the CONCEDE project which aims to harmonise dictionary encoding for several languages. Standardisation is still an important issue, the latest developments in SGML and XML methods are also discussed in some papers.

Besides the speakers from Central and Eastern Europe we have received and included some papers from the Far and Middle East countries as well.

The papers were selected by the international program committee:
József Andor, Christiane Fellbaum, Maurice Gross, Thierry Fontenelle, Ulrich Heid, Ferenc Kiefer, Júlia Pajzs, Max Silberztein, Tamás Váradi. We are very grateful for their help.

Júlia Pajzs

EuroWordNet: Results from the Italian Perspective

FRANCESCA BERTEGNA – ADRIANA ROVENTINI

Abstract

EuroWordNet (LE4003) is a project in the frame of the EC Language Engineering programme which emulates the work of George Miller and his group at Princeton University (Miller et al: 1990). The aim of the project was to build a multilingual semantic database in which different monolingual wordnets for European languages (in the first phase: Dutch, English, Spanish and Italian, and in the second phase, known as EuroWordNet2: French, Estonian, German, Czech) were linked by an Inter-Lingual-Index or ILI.

The paper is structured in two main sections: in the first section methods and goals of the project are introduced and, in particular, the methodology used to develop the Italian wordnet and the difficulties we met are described, both from a monolingual point of view and under a multilingual perspective. The second and main section of the paper is devoted to the detailed description of our semantic net. In particular, we give an overview of the vocabulary in our database, of its richness in terms of number of entries codified, variety and quantity of monolingual semantic relations. Significative taxonomies branches of both concrete and abstract nouns are illustrated to show how the taxonomic chains can be followed upwards to the highest concepts of our language and to the Top Concepts of the language independent hierarchy of concepts represented by the EuroWordNet Top Ontology. The structure of the semantic hierarchies and some of the problems we encountered when dealing with abstract and concrete nouns are reported.[1]

[1] Adriana Roventini wrote the first section of the paper (pp. 1-3) Francesca Bertagna the second one (pp. 3-11).

1. The building methodology and related problems

As is known, the original WordNet was built from scratch at the Cognitive Science Laboratory of Princeton University on the basis of psycolinguistic theories regarding human lexical memory. In this semantic net nouns, verbs, adjectives and adverbs are organised into synonyms sets each representing one underlying lexical concept. When constructing the Italian component within the EuroWordNet project we have taken the WordNet model as reference point and our core relation is also based on the synset, but we have extended the range of the different types of semantic relations, in particular to include cross-part-of-speech semantic connections, in the convinction that it would have been useful that a base concept such as *atto* (act), would be related not only to its synonym *azione* (action) and to the set of its hyponyms, but also to the near-synonym verb *agire* (to act), and, in the same way, the noun *attività* (activity) would be connected with its near-synonym adjective *attivo* (active). By means of all these relations, the word-meaning, seen and described from a multiple perspective, can be recognised and identified in many different contextualizations, and this feature appeared useful for future information retrieval applications. For a complete description of the principal internal semantic relations see (Climent et al: 1996 and Alonge: 1996).

Furthermore we differentiated our work from that of Princeton as far as the way of construction is concerned. In fact we have reused already existing lexical resources exploiting the results of previous projects (such as Acquilex and Delis) in which dictionary entries, in particular the definitions, were analyzed in order to identify different kinds of semantic relations and roles between word senses, e.g. hyperonymy/hyponymy, causative inchoative alternation, agent of, location etc..

The starting point to construct the multilingual database was the selection of a common set of Base Concepts agreed on by the EuroWordNet members to ensure an adequate coverage of the lexicon and an high degree of compatibility between the different monolingual databases. For Italian we selected this first core of lexical concepts from our monolingual database, LDB, on the basis of the number of their hyponyms (the criterion being "those most frequently used to define other words in dictionaries"). Further integrations were made from other sources (i.e. the Italian Reference Corpus) and by means of repeted comparisons and subsequent acquisitions of those base concept senses chosen by the other partners which had not emerged from the analysis of our data. In this way the first core subset increased up to 1059 items. This subset was accurately manually mapped to WN 1.5 and synonyms of each concept, when possible, were associated. This task was carried out by means of automatic extraction procedures followed by careful manual revisions, keeping in mind the definition of "weak synonymy" adopted by the project entailing the interchangeability of two words in a given context. When revising these automatically created synsets we found that a sense shifting often occurs. This phenomenon is unavoidable and must be controlled. In most cases the synsets appeared too large and manual revision was necessary to cut these synonym groups according to more coherent boundaries. In fact, although we were favourable to synsets sufficiently large to make clearer a concept in all its meaning nuances we had to avoid imprecise or misleading synsets. Once the base concepts were linked to the ILI and restructured in synsets, we extracted top down the first level of hyponyms and began the true developing of the semantic net through the changing and the remodelling of our primitive data structures.

The major problems we encountered in developing our network on the basis of our source data regarded the typical incoherencies deriving from the lexicographic metalanguage such as: circularity, under and over sense differentiation, inconsistency in the hyperonym assignment, hyperonym disjunction or conjunction. Then the decision to start from our resources, together with undeniable benefits represented by the great number of semantic relations already encoded, also entailed a great

effort and much manual work to reorganise and restrucure our taxonomic data in a different and more complex object such a semantic net. To give just an example, the under differentiation of many word senses originated very flat taxonomies in which too different types of hyponyms were found together (e.g. musical instruments with measure instruments, atmospheric phenomena with social phenomena and so on). This leaded us to introduce intermediate nodes in the taxonomies, often expressed by multiwords, usually constituted by noun + modifier, which restrict and specify the underlying lexical domains such as *strumento musicale*, *fenomeno atmosferico*, *legame familiare* (musical instrument, atmospheric phenomenon, family tie) etc...

Another problem we had to deal with was the difficulty of automatically mapping our synsets to the WN 1.5 ones. For this task we developed two different procedures: in a first stage we used a semi-automatic procedure based on the assumption that matching words in equivalent semantic hierarchies in different languages should refer to equivalent senses. Thus the procedure, starting from the lexical/semantic taxonomies we had constructed for the Italian database, attempted to match them against equivalent taxonomies in WordNet 1.5. The semantic context provided by the taxonomies should have allowed us to recognise the right sense in the target language of the word-sense we were examining. Unfortunately this procedure turned out not to be very efficacious: as far as nouns are concerned only the average of 20% of the analyzed entries were successfully mapped with a eq_synonym or eq_near_synonym relation. Furthermore the effectivness of this mapping procedure varies with the different categories of nouns it analyzes: the results were acceptable in the case of concrete nouns (e.g. entries in Animal taxonomies, or in the hierarchies of the most common instruments, plants or vehicles) but very insufficient when dealing with the Second Order taxonomies. These results arised from the following problems: (i) the bilingual electronic dictionary does not provide a traslation of the Italian word because the meaning is too specific; (ii) there are too many multi-words in WordNet that do not have any correspondence in our bilingual dictionary; (iii) there are too many differencies of classification as far as Second Order taxonomies are concerned. This issue is very important given that, how we have seen, the procedure was based on the correspondencies between taxonomies. For these reasons, in the second procedure we developed, we extended the research within the entire WordNet 1.5 and we introduced a statistical evaluation, or score of confidence, of the mapping depending on the different "itinera" or paths covered to assign the link.

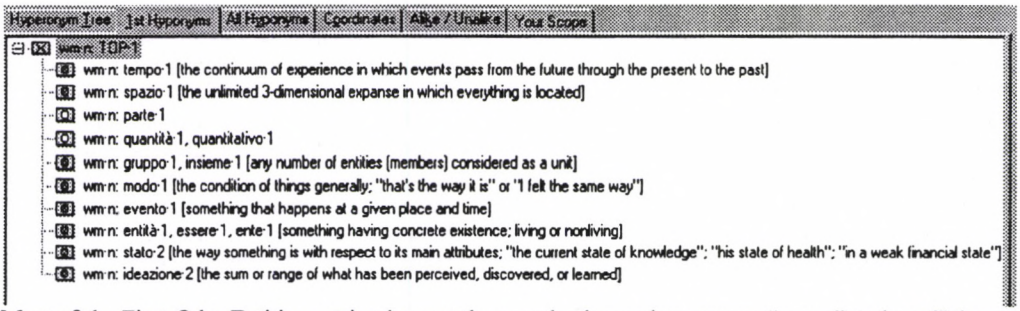
When the mapping is found in the same taxonomy of the Italian word-sense, together with other possible traslations in different WN taxonomies, only the mapping found in the same taxonomy is taken as good for the export file, while the other traslations are recorded in a separate file for eventual controls. As regards the score of confidence to be assigned when more than one mapping is found in the same taxonomy: a) if the mapping has as direct hyperonym the same direct hyperonym of the Italian word, the score is 100; b) in all the other cases, the score is 90; c) if the mapping is found in different taxonomies the maximum score, 80, is assigned when we find only one translation. For every additional traslation one point is subtracted to the score 80. Obviously the output of this second procedure had to be cheked as well, but the prominence that the score of confidence gave to the mappings reliability has been very useful to better address our revision work.

2. Results for Italian wordnet: some examples

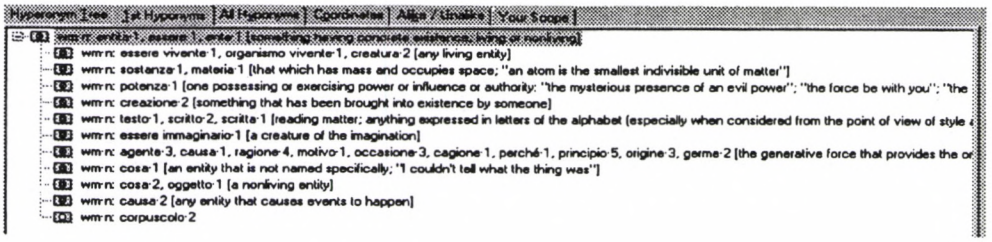
As above said, many internal relations were in our source but many others have been manually or semi-automatically added: this work concentrated on the Base Concepts subset but has been extended to many other synsets as well, with a good reliability. Given that the Italian wordnet has been built through successive top down extractions, at the very beginning, it was extraordinarely flat (2 or 3 levels). Today the work performed during EWN has leaded to a strong re-formulation of the trees

structures and the improvement in the taxonomies deepness makes the synsets more "meaningful", since they inherite information from all their hyperonyms. Our entries are now better organized and distributed in widely structured taxonomies. In carrying out this task we experienced that the codification and the reliability of the results varied depending on the belonging of the entry to the First or the Second Order nouns. In fact, while a correct redistribution of items in the concrete nouns taxonomies only implied much manual intervention, the abstract ones arised many more problems concerning the philosophy underlying the trees structures. Consequently the outline of a reference ontology required a lot of time and variuos changes up to the final design. Now all our entries are divided into three main divisions: First, Second and Third Order Entities.

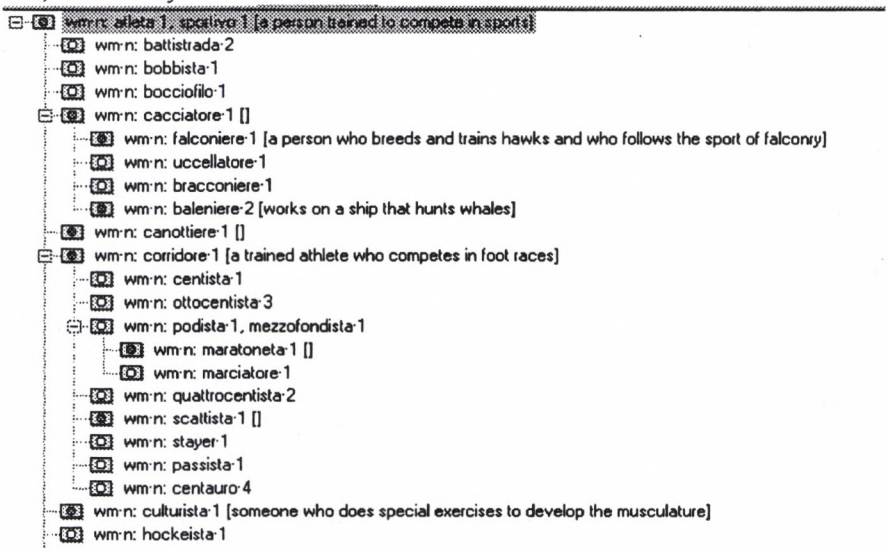
In the following section, we will illustrate the results achieved for some "basic" synsets belonging to the three Ontology subdivisions. First of all, we think that it can be useful to show the higher levels of the nouns taxonomies to analyze the tops. To do that, we will use a convention that will simplify the vision of the data: we will use a hypothetic "generative" node that we will call "Top": all the taxonomies derive from it. The picture below shows the immediate hyponyms of this node. As regards the First Order Entities, we can analyze the results starting from the more helpful base concepts, the ones with the higher number of hyponyms and relations:



Most of the First Oder Entities entries (except the ones having as hyperonym "group" and part") have been linked to the Top "*entità, essere, ente*", defined by the Italian dictionary as "*tutto ciò che esiste*" (everything existing). In our LDB no entry has as genus term the word *entità* (entity) and this could be seen as one more demonstration of the methodological and theoric problems raising from the bottom-up building strategy of our wordnet. The importance of the word meaning "entity" has been pointed out by the analysis of the lexical-gaps coming out from the comparison with the other languages: it has been possible, in this way, to choose this word-meaning in order to add an useful top level able to represent all concrete nouns of the First Order taxonomies (and it has been mapped to the WN1.5 synset "entity" defined as: everything existing, living or not living). In the following picture the most important Base Concepts in the First Order Entities are shown:

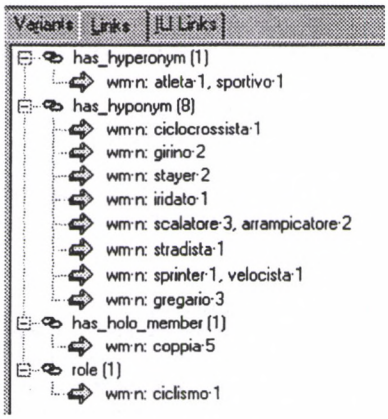


Considering the top-down structure, it is obvious that one of the nodes with the highest number of hyponyms is the one of the living entity: *essere vivente*. It is also linked to the collective noun *vita* (mapped to the synset life: living things as collective noun) by means of a holonymy relation and to the adjective *vivente* by means of a be_in_state relation. From this node three of the most important taxonomies of the Italian net derive:
{*persona, individuo, uomo, essere umano*} (human, mortal, someone, individual, person, soul)
{*animale, bestia, organismo animale*} (fauna, creature, brute, beast, animal, animate being)
{*pianta, organismo vegetale, vegetale*} (plant, flora, plant life)
The *persona* (human beings) taxonomy includes about 4200 synsets. In our source most of these entries were un-structured, most of them came from the noun *persona* (person) and from the pronoun *chi* (who), with this kind of definitory patterns:
macellaio (butcher): *chi vende carne* (someone selling meat)
poliziotto (policeman): *chi appartiene al corpo di polizia* (someone belonging to the police).
We tried to re-structure our entries, mainly manually or analyzing this definitory patterns and we succeed to distribute those about 4000 entries on different planes of the taxonomies. In the immediate hyponyms we find the numerous groups of professions (leaded by the synset *lavoratore* -a person who has an employment- and subvided in many groups: sellers, professionists, doctors, soldiers etc.), the group of followers of a doctrine or a school of thought, of the athlets, inhabitants, relatives, artists, intellectuals, experts and many others. When possible, we associated to all these entries a translation to the American-English of the ILI, when possible by means of an eq_synonym relation. When we could not find an eq_(near_)synonym, we used the eq_has_hyperonym relation. When possible, these main groups of hyponyms of first level have been internally structured.
Let's see, for example, part of the taxonomy of *sportivo, atleta* (athlet, joke), including, in all its levels, about 100 synsets:

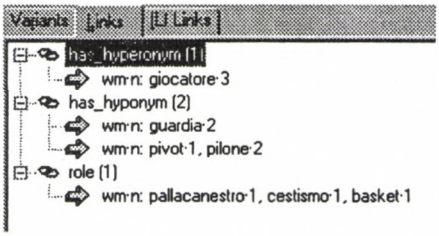


For some of these nouns we also codified relations to the practized sport, for example in the case of *cestista* or *ciclista*:

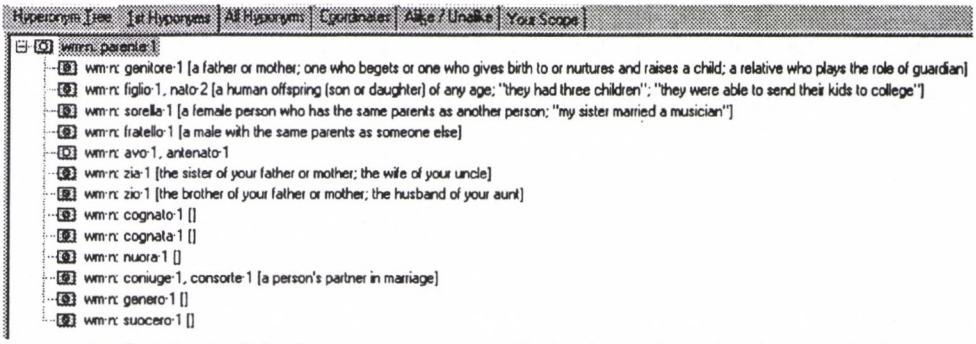
Ciclista (cyclist):



cestista (basket player)



Another important restructuration is concerned with the base concept of *parentela, legame familiare* (relationship, family tie) including 80 synsets, which are all referring to the base concept "relative".



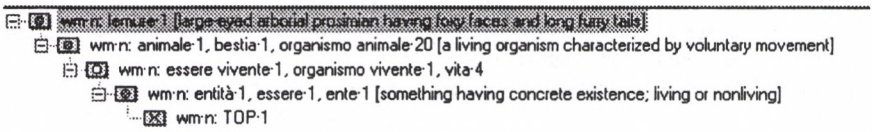
Now, on the first level of the Human taxonomy we find only 1500 entries, in general the ones for which we could not find a hyperonym able to subclassify the general term *persona*.

In these 1500 synsets we can also find all the nouns representing persons characterized only by some physical or moral qualities or by some kind of behaviour, such as: *bello* (nice, beautiful person), *brutto* (ugly person), *scemo* (silly person), *adulatore* (adulator) etc..

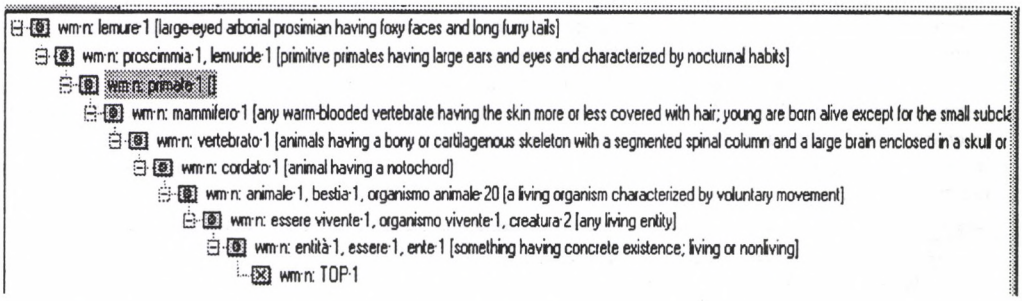
In many cases we could not find an equivalent synonymy in the ILI and a future project improvement could be to link all these nouns to all the correspondent adjectives, by means of an eq_be_in_state relation, to better specify the quality expressed by the noun.

Also in the Animal subset we had to better restructure too flat taxonomies directly derived from the lexicographic definitions. Most of the animals had the word *animale* as hyperonym, often followed, in the "differentia" part of the definition, by a specification of their biological-taxonomical family. For example, the synset *lemure* (lemur) was defined as follow:

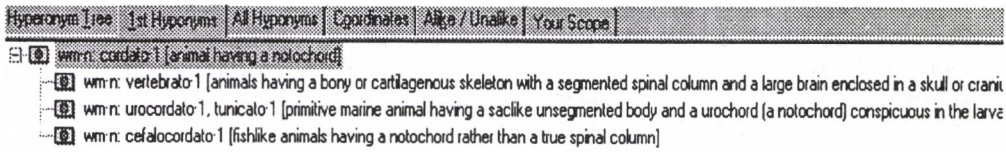
lemure: animale appartenente alla famiglia dei lemuridi (lemur: animal belonging to the Lemuride family). At the beginning our taxonomy was structured as follow:



After the analysis of the definition, the hierarchy earned many more levels (5) and now it appears as follow:

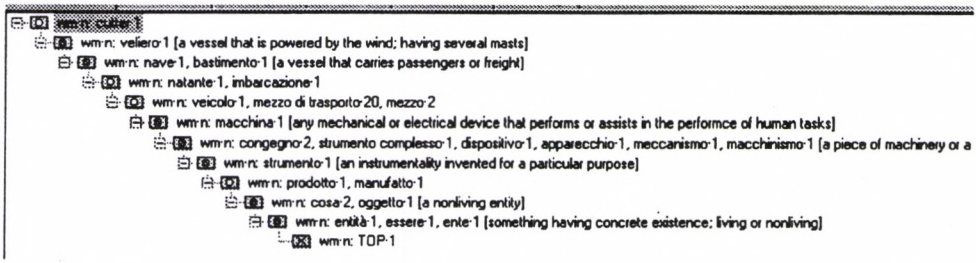


Some very important groups under a "natural" perspective were not so important under a lexicographic point of view and they had just a few or zero hyponyms.
That is true, for example, for the chordates, that groups all the vertebrate and others families.
The vertebrates, indeed, are not defined as chordate by the dictionary, but simply as animals and there is no definitions in our source that presents *cordato* as genus term .
Only the careful analysis of the Base Concepts selectioned by the other partners pointed out this taxonomically fundamental node. The definition
cordato: animale munito di notocorda: vi appartengono tunicati, vertebrati e cefalocordati
has been analyzed and we decided to put *cordato* over the subsets of vertebrate, tunicate and cephalochordate, as shown below:

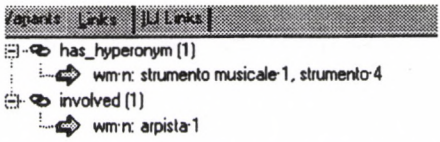


Today a big part of the whole animal subset (subset including about 800 synsets) has been re-classified on the basis of the analysis of the definitions and the definitive taxonomy is one of the deepest of the whole net, with branches of ten levels. Also in this large category we can add further information, adding relations to the verbs expressing the animal sounds or to the nouns meaning the different employments the humans make of them. We have already codified many of these relations, but not systematically yet. We are not close to the richness of WordNet1.5 that have so many synsets for the animals subset and a huge quantity of scientific terms and local variants, but we have to bear in mind that a very general lexicon has been codified in EuroWordNet (except for the part of Computer Terminology). On the basis of the use requirements we can think about further developments in the

taxonomies the user could need more, enriching, with a few efforts, the semantic relations already existing in the source or incrementing the number of the synsets using more specific sources.
That is all for what regards the living entity subsets (we didn not pay much attention to the plant subset consisting of about 550 synsets), but we should not forget to mention the little taxonomy deriving from living entity: the imaginary beings and creatures (35 synsets).
Many concepts in the First Order are not living entities, as, for example, the substances, the objects and the parts. Amongst the objects we can find many subdivisions, very important for the number of hyponyms and relations, like *mobili* (furnitures), *prodotti* (products), everything created by men with some purposes, *vestiti* (garments), many instruments such as *strumenti, arnesi ed attrezzi, strumenti musicali e di misura, meccanismi, veicoli* ecc..- (tools, musical instruments, measure instruments, devices, vehicles ecc..). We can try to give an overview of the most important and representative taxonomies we structured and the results we obteined.
One of the subset where we reached the highest number of level per taxonomies is the vehicles one, today consisting of about 350 synsets, structured and subdivided in main nodes like *veicolo a motore* (motor vehicle), *natante* (vessel, craft), *velivolo* (aircraft).
In the following picture we can see the taxonomical tree for cutter, a kind of vessel, in which we have reached 11 levels:

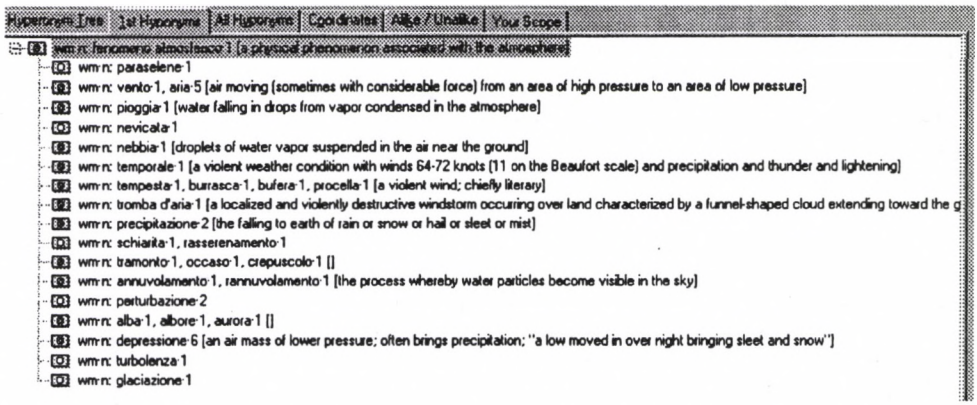


The hierachy of vehicle is under the more general concept of *strumento* (instrument); this taxonomy has many synsets for which we have encoded many telic relation (role/involved_instrument) in order to define to which kind of use the instruments are for.
We have also to underline that the hierarchical structure allows to inherit the semantic information top-down, from the highest concepts to the most specific ones by means of the hyperonymy relation: that means that if *pianoforte* (piano) is a musical instruments and for the "musical instrument" synset we codified the role_instrument relation to *musica* (music) and *suonare* (to play), pianoforte inherits the two relations from its hyperonym as well.
For musical instrument we encoded many semantic relations and they are inherited by all its hyponyms: Also for the instrument subset we can say that the original taxonomy we derived from the source was quite flat, but we have tried to redevelop it creating two new nodes for musical and measure instruments, each one having about 100 synsets. For the musical instruments we could add some intermediate levels, for example trying to group all the *strumenti a corda* (string instruments) and all the *strumenti a fiato* (wind instruments). At the moment, we have codified some relations to person who plays this specific kind of instrument, like in the case of *arpa* (harp) and *arpista* (harpist):

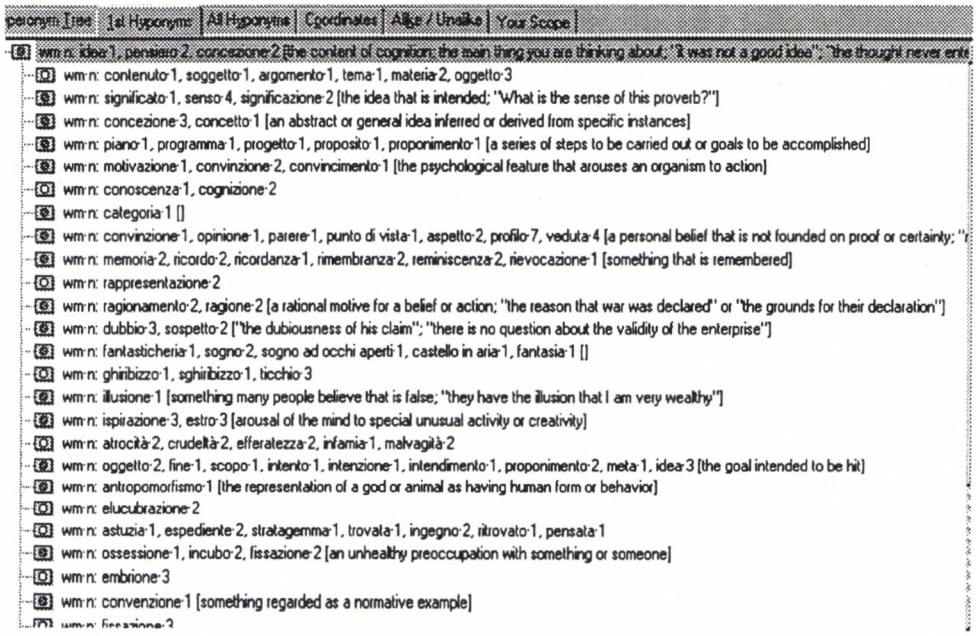


Obviously, starting from *arpa* we can go, via its link, to *arpista* (arpist) and all its taxonomy.

As far as the Second Order is concerned, the main base concepts by means of which the most part of the abstract nouns is organised are: event, state, way, time, quantity.
We show here below one of the subtaxonomies belonging to the base concept *fenomeno* (phenomenon) which has been restructured according to more coherent groups of concepts by introducing a few multiwords as intermediate nodes able to gather strictly related phenomena either natural or social. In their turn, natural ones branch out to atmospherical, geological, physical etc..
We can see the first level of hyponyms (18 synsets) of atmospherical phenomenon including 70 hyponyms distributed at various levels.



As a last example, we show the first level (32 synsets) of the Third Order Entities which is constituted by 1008 synsets, connected one to each other at different specifity levels. To this order belong, citing the definition in our Top Ontology, any kind of "unobservable proposition which exists independently of time and space, which can be true or false rather than real. They can be asserted or denied, remembered or forgotten. E.g. idea, though, information, theory, plan."



References

Climent, S., Rodríguez H., Gonzalo J. (1996). "Definition of the links and subsetsfor nouns of the EuroWordNet project", EuroWordNet Project LE4003, Deliverable D005. Http://www.let.uva.nl/~ewn.

Alonge, A. (1996). "Definition of the links and subsets for verbs", EuroWordNet Project LE4003, Deliverable D006. Http://www.let.uva.nl/~ewn.

Nancy Ide and Daniel Greenstein: Special Issue on EuroWordNet, Guest Editor: Piek Vossen, Volume 32, Nos. 2-3, (1998).

Miller G.A, Beckwidth R., Fellbaum C., Gross D., and Miller K.J. (1990). "Introduction to WordNet: An On- line Lexical Database", in: International Journal of Lexicography, Vol 3, No.4 (1990), 235-244.

Miller G.A. (1990). "Nouns in WordNet: a Lexical Inheritance System", in: International Journal of Lexicography, Vol 3, No.4 (1990), 245-263.

XML – A Solution for LDBs, EDs and MRDs?

MARIE–HÉLÈNE CORRÉARD – MATHIEU MANGEOT-LEREBOURS

Lexical resources are a key element of NLP applications. They come from different sources and in different formats. Users of lexical resources have to produce their own or process the available ones in order to make them compatible with their own environments and applications. The use of DML (Dictionary Markup Language), could make working with lexical resources easier. The nature of the resources available will be briefly examined, then the solution that adopted to "unify" them will be presented with concrete examples to illustrate the approach.

Introduction

Dictionaries and other lexical resources are a key element of NLP (natural language processing) applications. Often they come from different sources and in different formats.

Currently, users of lexical resources must either write their own dictionaries, not a trivial task, or process them to make them compatible with their own environments and applications.

This paper describes how the use of DML could simplify working with lexical resources, make possible their reuse and improve their shareability. Firstly, the available resources will be briefly examined, secondly the solution that was adopted to "unify" these resources will be shown, including a look at constraints and requirements and thirdly concrete examples will be presented before concluding on the future of such an approach.

1. Starting point

Dictionaries used in NLP vary greatly, according to the final purpose of the application they are used by, but essentially they are of two kinds: either they were designed specifically for computer applications or they were written for human users. In this paper the word dictionary is used to refer to ordinary, general language dictionaries created for humans whose texts exist in electronic format.

First, a quick survey was performed to find out more about what kind of lexical resources were needed and how the available ones could be improved. Accessibility of the data was one of the items that came high on the list of possible improvements. One way to make the data more accessible without altering the contents was 'standardisation' of the format. Then, several dictionaries were used to test this approach.

A brief description of each one of them is given below.

The Oxford-Hachette French Dictionary (OHD) is a bilingual dictionary. It consists of two sections roughly equal in size: French-English and an English-French. The dictionary is encoded in SGML. Its structure is fairly complex; a great number of elements are embedded.

The New Oxford Dictionary of English is a new monolingual dictionary. It contains most of the elements of a monolingual dictionary, including etymology, sample material and encyclopedic information.

The Password semi-bilingual English-French dictionary consists of one developed semi-bilingual section and one French index which cross-refers to English entries in which the French word is given as a translation.

The FeM (French-English-Malay) dictionary supplies English and Malay translations of the French entry. English was used as a help for lexicographers during the dictionary development.

2. Solution

In order to get around the difficult it was decided to adopt, at a higher level, a common format for all the dictionaries. This common standard format had to be easily readable and to make it possible to keep all the information which was present in the original format. Then tools based on this common format could be built.

This section describes the requirements and gives an explanation of how the common format was defined.

2.1. Requirements

The design of the solution was driven by a list of specifications. These specifications came from previous experiments in computational lexicography and lexicology such as the indexation of the French-English-Malay dictionary [Lafourcade96], the building of the French-UNL database [Mangeot97], [Mangeot98] or the computerization of the I. Mel•uk's Combinatory and Explanatory Dictionary of contemporary French [Sérasset98].

It was essential to find a way to preserve all the information present in the original format of the dictionary during the conversion. The dictionaries might be used for various applications, so it was not possible to predict in advance the kind of information that should be kept or left out.

In order to guarantee a maximum of compatibility for the new format and to reuse previous work in the domain, the obvious approach was to use existing norms and standards as much as possible. Furthermore since most of the resources available at the time were encoded in SGML [ISO86] it seemed reasonable to try and chose a format which did not need a lot of conversion work.

On the one hand, the power of object programming as well as that of relational database query facilities were attractive. On the other hand, the opacity of data repositories and portability problems were decisive factors for the choice of a textual format, either for storage or exchange when manipulating dictionaries.

2.2. Format adopted

All these considerations led to the choice of eXtended Markup Language (XML) [Connoly97] for encoding the dictionaries. XML is a W3C recommendation [W3C98a]. It is also UNICODE [ISO93] compliant. XML makes it possible to represent a large variety of information. All these features guarantee readability, perenniality and compatibility with an increasing number of tools.

Furthermore, because XML is a subset of SGML, the conversion of SGML dictionaries, well formed according to XML, into XML is unnecessary. Also, XML is a textual format, therefore it

will always be possible to read the original files encoded in XML.

Now that the format is defined, a problem remains: how to encode the structure of the dictionaries? Two alternatives are possible:

Using a general DTD

The first option was to define a general DTD. This DTD would have to be generic enough to allow the encoding of all the dictionaries currently available. The conception of tools would then be easy because all of them would be based on the same DTD. This solution, despite its simplicity, was rejected because it was not possible to convert all the dictionaries following the same DTD without loss of information. It was also obvious that each dictionary has its own particular structure and, except for some rare cases, it was impossible to convert all the contents of one dictionary into another dictionary structure.

Keeping the original structure

An easier solution was to keep the original structure of each dictionary. A difficulty then rises at the stage of designing a tool for more than one dictionary. It appears quickly that each dictionary requires its specific tool. Therefore this solution does not solve all the problems.

A hybrid solution

A hybrid solution was then envisaged. XML is designed to be used with namespaces [W3C99]. It seemed appropriate to introduce a new one, specialized for dictionaries: DML for Dictionary Markup Language. This namespace is used for a hierarchised restricted set of tags. This set is composed of tags describing the same information in different dictionaries. For example, <dml:entry> always refers to an entry or <dml:headword> to the headword of an entry.

When some information in a dictionary cannot be represented with a tag from the DML set, it is still possible to copy it from the source file without transforming it. Specific tools manage it as they would the original file. If this type of information is present across several dictionaries, a new tag is then added to the DML set. The DML tags are used by the various tools as points of reference in an unknown converted dictionary.

The set of tags is composed of tags coming from standards like TEI/MARTIF [Ide95], [Johnson95], [Melby94], [ISO95]; GENELEX/EAGLES [GENELEX93] and GENETER [GENETER98]. The matching between a DML tag and an original tag is performed by a linguist to avoid possible conflicts between the tags.

Here is an alpha version of the DML tagset. The tags were chosen on the basis of their frequency. If an element occurred in more than 2 dictionaries (this figure may change at a later stage) it was added to the tagset. The tagset itself is evolving as new dictionaries are explored and converted.

```
<dml tag>                                          (tei equivalent)

<dictionary
    name=""
    date=""
    source-language=""
    target-language="">
<letterset letter="">
<entry>                                            (entry)
    <headword homograph-number="">                 (hom)(orth)
    <headword-variant>                             (oVar)
```

<pronunciation>	(pron)
<phonetic encoding="">	
<etymology>	(etym)
<syntactic-cat>	(sense level="1")
<part-of-speech>	(pos)(subc)
<semantic-cat>	(sense level="2")
<indicator>	(usg)
<label>	(lbl)
<definition>	(def)
<example>	(eg)
<translation language="">	(trans)(tr)
<collocate>	(colloc)
<xref>	(xr)
<x-headword homograph-number="">	
<x-syntactic-cat>	
<note>	(note)

The next section shows how conversion was performed using DML tagset and how the results were exploited.

3. Examples

3.1. Conversion

According to the source format of the dictionary, there are three types of conversion. The simplest type occurs when the source format is well-formed SGML; the second type, when all the information is under the form attribute-value, and the third, the most complex one, relates to typographic formats which have to be parsed in order to extract as much information as possible.

3.1.1. Well-Formed SGML

If the dictionary is encoded in SGML and "well formed" in the XML sense (ie all opening tags are closed and the file is parsable by a context-free grammar), the conversion is very easy, since the structure is already, de facto, in XML. In this case, the only tasks are: conversion of characters into UNICODE characters set, changing the file encoding to UTF-8 and adding as much DML tags as possible.

If some information is redundant between the DML tag and the original tag, the latter is replaced and a note is kept of the changes. If the replacement DML tag is less precise than the original one, the original one remains in the text, embedded inside the DML tag. If some information is not in the same format (eg an element instead of attribute), it is altered to conform to DML and a note is kept of the changes.

The example is taken from the OHD [OUP-H94].

First, here is a sample of the entry abréger in original format:

```
<se><hw>abr&ea.ger</hw><pr><ph>abKeZe</ph></pr><hg><ps>vtr</ps></h
g><s2 num=1>(<ic>rendre court</ic>) to shorten
[<co>mot,expression</co>]; to summarize [<co> texte,
discours</co>]; <sl>&hw. &oq.t&ea.l&ea.vision&cq. en
&oq.t&ea.l&ea.&cq.</sl> to shorten &oq.television&cq. to
&oq.TV&cq; (...) </se>
```

The headword abréger is followed by its pronunciation in Alvey notation, its part of speech vtr and its English translation to shorten then to summarize; the translations are differentiated by context (collocates). An example follows: abréger 'télévision' en 'télé' then its translation: to shorten 'television' to 'TV'. Translations were left

untagged.

The sample below is the same entry with DML tags. Modified parts are in italics

```
<dml:entry><dml:headword>abr&#xE9;ger</dml:headword>
<dml:pronunciation><dml:phonetic encoding="ALVEY">
abKeZe</dml:phonetic></dml:pronunciation><hg><dml:part-of-
speech>vtr</dml:part-of-speech></hg><dml:semantic-sense>
<ic>rendre court</ic> to shorten <co>mot, expression</co>; to
summarize <co>texte, discours</co>; <sl>&hw;
&oq;t&#xE9;l&#xE9;vision&cq; en&oq;t&#x E9;l&#xE9;&cq;</sl> to
shorten &oq;television&cq; to &oq;TV&cq;;</dml:semantic-
sense></dml:entry>
```

3.1.2. Attribute-Value

When the original dictionary is represented by series of attribute-value pairs, the conversion remains simple. It consists in devising a DTD for the dictionary and converting the attribute-value pairs into <tag>value</tag>. Characters and file encoding are also converted.

The example for abréger below is taken from the FEM [Lafourcade96].

```
(:fem-entry
(:ENTRY "abréger")
(:FRENCH_PRON "abre-je-")
(:FRENCH_CAT "v.tr.")
(:FRENCH_GLOSS "un texte")
(:ENGLISH_EQU "to shorten")
(:ENGLISH_EQU "to abridge")
(:MALAY_EQU "memendekkan")
(:MALAY_EQU "meringkaskan")
)
```

The entry after conversion looks as follows:

```
<dml:entry><dml:headword>abr&#xE9;ger</dml:headword>
<dml:pronunciation><dml:phonetic encoding="GETA">abre-je-
</dml:phonetic></dml:pronunciation>
<dml:part-of-speech>v.tr.</dml:part-of-speech>
<FRENCH_GLOSS>un texte</FRENCH_GLOSS>
<dml:translation language="en">to shorten</dml:translation>
<dml:translation language="en">to abridge</dml:translation>
<dml:translation language="ml">memendekkan</dml:translation>
<dml:translation language="ml">meringkaskan</dml:translation>
</dml:entry>
```

3.1.3. Typographic Format

The most complex case occurs when a dictionary needs to be converted from a typographic format such as typesetters' tape, word processor, HyperText Markup Language (HTML). One particularly complex aspect these formats is that they represent knowledge designed to be readable by humans who can infer structure and disambiguate senses easily. In order to extract the information and, above all, build a deep structure for such a dictionary, a powerful tool built by [Haï98] called RECUPDIC was used. This tool combines two methods: a string transducer and a special tree parser. The structure of the result is described as a grammar and the tool extracts as much information as possible.

Here is the entry babble from Password semi-bilingual English-French dictionary:

```
>U43<babble >U1<[\B.270babl] >U2<verb >U23<1\N>U1<to talk
indistinctly or foolishly: >U2<What are you babbling about now?
>U8< bafouiller, bavarder\L
>U23<2\N>U1<to make a continuous and indistinct noise: >U2<The
stream babbled over the pebbles.>f5h8<. >U8< gazouiller\L
```

and the same entry converted in XML:

```
<dml:entry><dml:headword>babble</dml:headword>
<dml:pronunciation><dml:phonetic
encoding="Password">'babl</dml:phonetic ></dml:pronunciation>
<dml:syntactic-cat><dml:part-of-speech>verb</dml:part-of-speech>
<dml:semantic-cat num="1"><dml:definition>to talk indistinctly or
foolishly </dml:definition><dml:example>What are you babbling
about now?</dml:example>
<dml:translation>bafouiller</dml:translation>
<dml:translation>bavarder</dml:translation><dml:semantic-cat
num="2"><dml:definition>to make a continuous and indistinct
noise</dml:definition><dml:example>The stream babbled over the peb
bles.</dml:example><dml:translation>gazouiller</dml:translation></
dml:semantic-cat>
</dml:syntactic-cat></dml:entry>
```

A summary of conversion operations is described in the table below:

Dictionary	Format	Size (in bytes)	Time spent
OHD - en/fr	SGML	17 Mb	1 day
OHD - fr/en	SGML	15 Mb	1/2 day
NODE - en	SGML	38 Mb	1 day
Password - en/fr (letterset)	typesetter's tape	300 Kb	5 days
Password - en/ja (letterset)	typesetter's tape	250 Kb	1 days
FeM - fr/en/ml	attribute-value	9 Mb	1/2 day

3.2. Usage

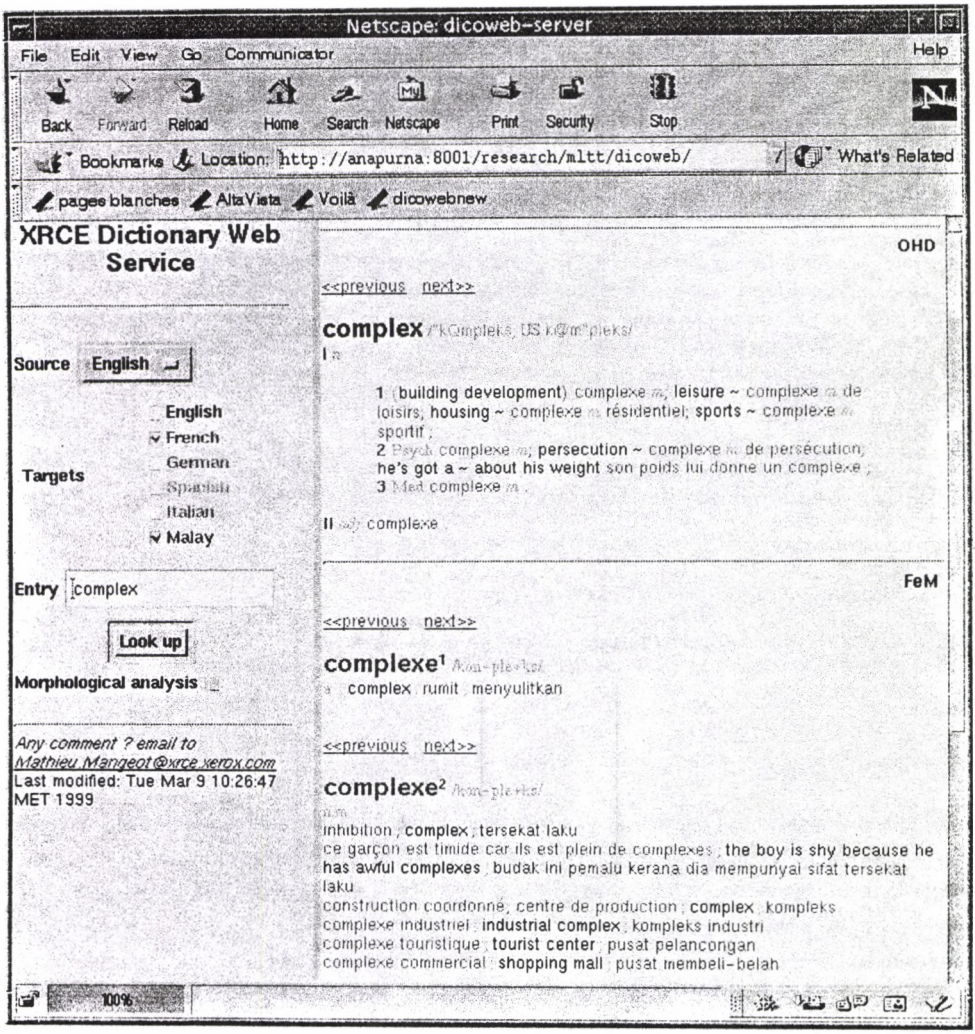
Some dictionaries do not contain information corresponding to some of these DML tags and some others contain information that is not covered by the DML tagset. However, as the tools are based on the DML tagset they will always find those elements which are represented by the tagset and present in a given dictionary, eg the <dml:headword> tag will always refer to the headword of an entry. Tools must be evolutive, to take into account the changes of DML.

As the resources are encoded in XML, all XML-compliant tools can be used. For example, a dictionary can be exported into a specific format with the help of XSL [W3C98b] or DSSSL [ISO96]. Tree transformations operations become possible. Dictionary readability can be improved with an associated stylesheet and an XML-compliant browser. Because of the relative youth of XML, few good tools are available yet but there should be more in the near future.

Two applications realised with XML/DML-encoded dictionaries are presented below.

3.2.1. Dicoweb

Dicoweb is a dictionary webserver. It was designed for human usage. It is used for experiments and research[1]. For legal reasons, not all these dictionaries are accessible to the public.



The Dicoweb user first selects the source language of the headword she is looking up, then she selects the target language(s). The user can select as many target languages as are available. Before consulting the dictionaries, she can process the headword through a morphological analyser. Two buttons, labelled "previous" and ""next", give access to the preceding and following entries in dictionary order. For clarity reasons, each language is visualised in a specific colour and font.

[1] URI: http://silfide.imag.fr

A Common Gateway Interface (CGI) script written in PERL [Wall91] works as a link between the user, the morphological analysers and the dictionaries. Dictionaries are selected according to the languages checked by the user. The files corresponding to the dictionaries are browsed by the script looking for a PERL regular expression such as:
`"/<dml:headword[^>]*>$ENTRY<\/dml:headword>/"`
where `$ENTRY` represents the headword entered by the user.

XML browsers are not widespread so it was decided to convert the result into HTML before sending it back to the user. The pages are built on the fly, with no breach of copyright and the possibility to modify directly the rendering of the final page.

Adding a new resource simply means adding the location of its file and the languages it covers to the script.

3.2.2. XeLDA

XeLDA (Xerox Linguistic Development Architecture) was built to provide developers and researchers with a common development architecture for the open and seamless integration of linguistic services. These services may include such applications as translation aids, syntax checking, terminology extraction, and authoring tools in general.

The above sample of Password English-French semi-bilingual dictionary was integrated into XeLDA. The dictionary was converted to comply with XeLDA DTD with the help of an XSL stylesheet. Here is the result of the transformation:

```
<xbdict>
<entry>
<headword><spl>babble</spl></headword>
<hwinfo><pronunciation><phonetic>['babl]</phonetic>
</pronunciation></hwinfo>
<syntactic><senseinfo><pos>verb</pos></senseinfo>
<semantic>
<subsense>to talk indistinctly or foolishly: What are you babbling
about now?</subsense>
<subsense><trans>bafouiller</trans></subsense>
<subsense><trans>bavarder</trans></subsense></semantic>
</syntactic>
</entry>
```

Conclusion

The work presented in this paper is still under development. The current results are satisfactory. However, further investigation is needed to establish the adaptability and coverage of DML. In the longer term it is planned to build new tools that will enable a user to set personal parameters according to the task at hand.

References

[Atkins94] **Atkins, B. T. S. and Zampolli, A. (1994)** *Computational Approaches to the Lexicon.* Oxford University Press, 480 p.

[Bauer94] **Bauer D., Segond F. & Zaenen A. (1994)** *Enriching a SGML-tagged Bilingual Dictionary for Machine-Aided Comprehension.* Technical Report, Xerox Research Center Europe, 21 p.

[Boguraev89] **Boguraev et al. (1989)** *Computational lexicography for natural language processing.* B. Boguraev & T. Briscoe, ed., Longman, London & New York, 310 p.

[Briscoe93] **Briscoe et al. (1993)** *Inheritance, Defaults and the Lexicon.* Cambridge University Press, Cambridge.

[Connoly97] **Connolly, D. (1997)** *XML Principles, Tools and Techniques* World Wide Web Journal, Volume 2, Issue 4, Fall 1997, O'REILLY & Associates, 250 p.

[Fedder91] **Fedder et al. (1991)** *Typed Feature Logic and its role in MULTILEX.* Centre for Computational Linguistics, UMIST, november 1991, 30 p.

[GENELEX93] **GENELEX (1993)** *Projet Eureka Genelex, modèle sémantique.* Rapport Technique, Projet Eureka,

Genelex, 4 march 1994, 185 p.
[GENETER98] **GENETER (1998)** *Modèle générique de représentation des données terminologiques.* URI: http://www.uhb.fr/Langues/Craie/balneo/nen3.zip.
[Haï98] **Haï, D. (1998)** *Techniques génériques d'accumulation d'ensembles lexicaux structurés à‡ partir de ressources dictionnairiques informatisées multilingues hétérogènes.* Thèse de nouveau doctorat, Spécialité Informatique, Institut National Polytechnique de Grenoble, 168 p.
[Heid92] **Heid et al. (1992)** *Extracting linguistic information from machine-readable versions of traditional dictionaries, a metalexicographic method and some tools.* Proc. COMPLEX'92, Conference on Computational Lexicography an Text Research, Budapest, Hongrie, Linguistics Institute, Hungarian Academy of Sciences, Budapest, pp 161-174.
[Ide95] **Ide, N. and Veronis, J. (1995)** *Text Encoding Initiative, background and context.* Kluwer Academic Publishers, 242 p.
[ISO86] **ISO (1986)** *IISO 8879 (SGML) Information processing -- Text and office systems -- Standard Generalized Markup Language*, Geneva, 155 p.
[ISO93] **ISO (1993)** *ISO/IEC 10646 (UNICODE) Information technology -- Universal Multiple-Octet Coded Character Set (UCS)*, Geneva, 754 p.
[ISO95] **ISO (1995)** *ISO DIS 12620 (MARTIF) Terminology - Computer Applications– Data Categories.* ISO TC 37/SC 3/WG I, Geneva.
[ISO96] **ISO (1996)** *IISO/IEC 10179 (DSSSL) Information technology -- Processing languages -- Document Style Semantics and Specification Language*, Geneva, 292 p.
[Johnson95] **Johnson, E. (1995)** *The Text Encoding Initiative.* TEXT Technology vol. 5, n°3, Autumn 1995, pp 174-175.
[Lafourcade96] **Lafourcade M. (1996)** *Structured Lexical data: how to make them widely available, useful and reasonable protected? - a practical example with a trilingual dictionary.* Proc. COLING-96, Copenhagen, Denmark, Vol 2/2, pp. 1106-1110.
[Mangeot97] **Mangeot-Lerebours, M. (1997)** *Outils pour lexicographes naïfs (en informatique).* DEA Informatique Systèmes et Communications, GETA-CLIPS-IMAG, Université Joseph Fourier Grenoble 1, 19/06/97, 58 p.
[Mangeot98] **Mangeot-Lerebours M. (1998)** *Conception, implémentation et indexation de BaLeM, une base lexicale multilingue.* Proc. TALN'98, Paris, pp 215-217.
[Melby94] **Melby et al. (1996)** *The Machine Readable Terminology Interchange Format (MARTIF), Putting Complexity in Perspective,* Termnet News, vol 54/55.
[OUP-H94] **Oxford-Hachette (1994)** *Le dictionnaire Hachette-Oxford* Oxford University Press & Hachette, 1950 p.
[Sérasset98] **Sérasset, G. & Mangeot-Lerebours M. (1998)** *L'édition lexicographique dans un système générique de gestion de bases lexicales.* NLP-IA'98 Moncton, NB, Canada, vol 1/2, pp 110-116.
[Thurmair98] **Thurmair, Gr. et al. (1998)** *The Open Lexicon Interchange Format (OLIF).*
URI: http://www.otelo.lu/seite2.htm
[W3C98a] **W3C (1998)** *XML 1.0* URI: http://www.w3.org/TR/1998/REC-xml-19980210
[W3C98b] **W3C (1998)** *XSL 1.0* URI: http://www.w3.org/TR/1998/WD-xsl-19981216
[W3C99] **W3C (1999)** *XML namespaces*
URI: http://www.w3.org/TR/1999/REC-xml-names-19990114
[Wall91] **Wall L. & Schwartz R. L. (1991)** *Programming PERL,* O'Reilly and Associates.

The SGML/TEI Model and Its Application to the Encoding of Electronic Dictionaries

ANA FERNANDEZ-PAMPILLÓN – MARÍA MATESANZ DEL BARRIO –
ALFREDO FERNÁNDEZ VALMAYOR – COVADONGA LÓPEZ ALONZO

The main aim of the research described in this article is to facilitate the automatic extraction of lexical information contained in a traditional dictionary, through the explicit description of the structure of that information. For this purpose we are using a descriptive data model to the lexical information of the dictionaries entries. This model is based on a descriptive markup language, the TEI model for electronic dictionaries. This model is defined with the markup metalanguage, SGML, which is hardware and software independent. The aim of our work is to facilitate the automatic extraction of the information contained in an electronic dictionary, through the explicit description of the information structures, the relationships among them and their attributes. The explicit description of the dictionaries entries, provides the systematic and precise (rigorous) minimum requirements to be used directly both, by computer systems and human users. Furthermore, the advantage of explicitly defining the structure of the lexical information with respect to the traditional database approach, is greater flexibility in the design process of the dictionary. The SGML/TEI model has been applied to one of the most important general dictionaries in the Spanish language, the DRAE.

1. Introduction

Dictionaries have traditionally been the most important lexicographical works in the lexical description of a language and have been organised using lexicographic criteria, not always clearly defined. Until recently, dictionaries have used printed format and have been used only by reference language readers. Nowadays, new computer technology enables all lexical information to be stored in electronic format, such as CD-ROM or magnetic disc. This change to electronic format, or to be more precise, the change in the underlying technology (the incorporation of computer technology) brings about many advantages. These advantages include greater speed and flexibility for the user to access the sought after forms. However, the greatest advantages come from the use of computer technologies. These imply a clear definition of the data structure used to represent the lexical information of dictionaries. In the most simple of cases, the dictionary is conceived as a text , that is to say, the underlying data structure is only the lineal sequence of characters. However, in more

significant cases, both in theory and in practice, the dictionary is considered as a data base with lexical information. Thus, the underlying data structures are complex entities and allow another kind of relationship, as well as the purely sequential. These data structures are at the base of algorithms which allow optimum access to lexical information, control of data consistency and concurrent and secure access to the same data by various users. Another important consequence of the technological change arises if machine readable dictionaries (MRD) are organised as a database. These dictionaries can be used as a source of lexical knowledge for other computer systems and especially natural language processing systems (NLP).

Dictionaries have a series of features suitable for extracting lexical information, which has allowed the semiautomatic building of computational lexicons [Martí et al., 1998]. However, they do not comply with rigorous minimum requirements to be used directly by computer systems. Furthermore, dictionaries also have problems and errors of a lexicographic nature. These defects tend to accumulate in consecutive editions, are difficult to be eradicated without clearly defining a data model which allows the lexical information to be systemised.

The increasing importance of the lexical component in the NLP systems is a direct consequence of the increasing importance of linguistic theories of a lexical-based approach e.g. Lexical Functional Grammars and the corpora-based statistical approach. [Kaplan&Bresnam, 1982] [Boguraev&Pustejovky, 1996].

These PLN systems require a great quantity of lexical knowledge to effectively process the language [Guthrie, 1996]. In the last decade, great efforts have been made to find standard models to define lexical entries [Calzolari, 1994], and to define methodologies of automatic acquisition of lexical knowledge using corpora and dictionaries, that already exist in printed form [Varile&Zampolli, 1992] [AQUILEX, 1995][EAGLES, 1996]. Likewise, projects to develop new dictionaries or computational lexicons are underway, some with an automatic translation approach, such as the Japanese project EDR (Electronic Dictionary Research) [Yokoi, 1995] and some with a more generic approach such as the lexical database WordNet and EUROWordNet [Miller, 1993][Gonzalo et al., 1998], the Alvey Lexicon [Grover et al., 1993] and COMLEX, one of the lexicons created by the LDC (Linguistic Data Consortium) [Grishman et al., 1994]. Two fundamental problems arise from the building and sheer size of these new lexicons, to control the consistency of the data and the automatic or at least, semiautomatic insertion of new data [Castellon, 1992].

To control the consistency of the data stored in the lexicon implies basically solving the problem of the redundancy of data and, consequently, their maintenance and coherence. The manual acquisition of new lexical knowledge is both costly and inefficient, due to the volume of information. New knowledge from other sources of lexical knowledge such as textual corpora and already existing MRDs, must be inserted [Boguraev&Pustejovsky, 1996]. This last point implies, in some cases, the possibility of inserting the lexical knowledge contained in traditional dictionaries into the new computational lexicons, and in other cases converting the traditional dictionary into electronic format which could be used both by a human user and by a PLN system.

Regarding this subject, the main aim of the research described in this article is to facilitate the automatic extraction of lexical information contained in a traditional dictionary, through the explicit description of the structure of that information. A data-based model, as previously explained, can be used to reach this objective. The design of a database compels us to explicitly define the basic structures of the information structures comprised in the dictionary, its substructures and all relationships among them. This approach has obvious advantages. Firstly, the possibility of selecting the database model most suitable for our purpose (e.g. relational, network, hierarchical). Secondly, the possibility of designing a database with minimum redundancy, which would ensure a greater control of data consistency, and efficient and secure management of information using already existing powerful DataBase Management System (DBMS) [Elmasri, 1997].

However, we have not followed database approach in the work presented here. In our work, the electronic dictionary is conceived as a text, that is to say, the file structure is, simply, a lineal sequence of chars. The difference being, we include tags in this text (markup text), and we use

textual tags to explicitly define the dictionary lexical information. These tags can also include the attribute and the relationships that can exist among the structural elements.

The advantage of explicitly defining the structure of the dictionary, with respect to the database approach, through tags, is greater flexibility, both in the design process and in the maintenance of the dictionary. This is due to the possibility of making changes in the data scheme, at a much lower cost. When the data scheme is explicitly inserted into the data in the same file, the underlying scheme of the data (sequential text) does not change. However, with a database approach, the data scheme must be finished before building the database, and the modifications therein, could only be minimal. This is difficult in many cases, due to irregularities in the information structure, especially in the case of the lexical information of a traditional dictionary entry. This situation is clearly illustrated when describing our structure analysis of a traditional Spanish dictionary.

In this article we are going to present a model for the explicit description of lexical information of MRDs, based on a markup language, or to be more precise, the markup metalanguage description, SGML, and the TEI guidelines for electronic dictionary encoding. In point three, we will go on to describe the data structure of MRD entries and in particular of the *diccionario de la lengua española* of the Spanish Royal Academy [DRAE, 1992]. This dictionary is one of the most important general dictionaries in the Spanish language and one of the most complex. In point four, we will describe the application of the SGML/TEI model to the information of the DRAE entries. Likewise, we will present the results obtained and the methodology used. Finally, we will summarise the main results obtained and will point out the current work lines.

2. The SGML/TEI model

Descriptive markup is defined as "text that is added to the data of a document in order to convey information about it"[Goldfarb,1990]. SGML[1] (ISO8879) provides not only a set of standardised codes that add meta-information to an instance document. SGML provides a language that designers can use to formally define the structure of a whole class of documents. In SGML, this structure is formalised through a DTD (Document Type Definition) that declares the elements that compose the instance documents. A DTD is the grammar that formally describes the structure of a class of texts. A DTD describe the structure of documents in a way that is already familiar to linguists and designers of programming languages. It provides a notation to express a basic tree-like structure attaching a tag to each tree node. Tree nodes, tagged elements, can also be decorated with attributes expressing contextual relationships or any other kind of information we want to add to the document.

The SGML/TEI model is a set of general and descriptive tags defined with the standard metalanguage SGML. It is a general model since it is applicable to any type of text and to any language. As already pointed out, the basic aim of markup languages defined with SGML is the explicit description of its structure and properties of any type of data in a standard way and independent of the application or computer system that is to process the data. The advantages of using markup languages defined with SGML to describe information are (1) that this information is independent of the hardware/software platform and (2), suitable for electronic interchange. In SGML, the document structure and properties[2] are defined in a special document, called DTD. The different structural units comprising a document are called elements and are marked in the text between two tags, one at the beginning and the other optional one at the end. The elements are defined in the DTD by a generic identifier. Properties and information about their processing can be included using attributes.

Different languages are defined to describe the same information with SGML. If the information is intended to be reused, a common descriptive language must be found. With an aim to

[1] Standard Generalised Markup Language
[2] The document is the type of basic data in SGML

unifying formalisms, the Work Group of the international project[3] Text Encoding Initiative (TEI) has proposed a set of marks, unique for each type of textual information in electronic format [Sperberg-McQueen&Burnard, 1994].These include a model for MRDs.

There are several reasons for our using the TEI model for the description of lexical information of dictionaries. Firstly, being based on the SGML, the marked information is portable and reusable. Portability refers to the effort required to transfer the data from one hardware and/or software to another, and reuse is the capacity to use all the data or part of it. These two features are especially important in the case of electronic linguistic resources (dictionaries, corpus) which can be qualified as *quality* [Pressman, 1991]. Moreover, the software applications making use of this information will be independent of the hardware, the localisation and the set of tags describing the data [Sperberg-McQueen, 1995].

Secondly, the TEI model has another series of interesting properties: (1) It is suitable for the type of data (text) of sources of lexical knowledge (dictionaries, text corpus, etc.). (2) It is translinguistic (regardless of the language) and (3) multidisciplinary (applicable to texts from different disciplines and textual typologies). (4) It allows multiple-view definition, that is, the same information can be viewed in different ways. The TEI model proposes two views for MRDs, both editorial and lexical. The first describes the information for editing printed dictionaries. The second describes the entries in a way that can be used for analysis and linguistic investigation. (5) Finally, in the TEI model, the tags are organised in modules which can interrelate with each other in an organised and preestablished way. This modular structure facilitates modifications during the building and maintenance of the dictionary and allows the idiosyncratic features to be adapted to each language and textual typology.

However, as will be explained in point four, the application of the TEI model is not an easy task, since it is a very complex model. The reason being its generality. The TEI tries to describe the information of any dictionary and any language[4] [Ide&Veronis, 1995].The result is a wide ranging and irregular model. A set of 40 basic tags is suggested for describing the information of an entry, along with others common to any TEI document[5], and others which are optional. Furthermore, the lexical elements of an entry can be subelements to almost any other element of the entry.

The TEI basic tags for describing the lexical information of the entries are classified into two levels, according to their level of description: Top level tags and phrase level tags. The top level tags structure the entry information in almost hierarchical level and describe this information very generally. The phrase level tags describe the subelements that constitute the top level elements, and refine the description. Figure 2.1 shows the structure of the entries at top level defined by the TEI[6]. Figure 2 shows an example of how the structure of an entry is explicitly described with two meanings. Using the tags <*entry*> and <*sense*>.

```
SUPERENTRY ::= (FORM)?, (ENTRY)+
ENTRY ::= (HOM | SENSE | DEF | EG | ETYM | FORM | GRAMGRP| NOTE | RE | TRANS | USG |XR )+
ENTRYFREE ::= (TEXTO)+ | DEF | EG | ETYM | FORM | GRAMGRP| NOTE | RE | TRANS | USG |XR
              |PHRASE | INTER
HOM ::= SENSE | DEF | EG | ETYM | FORM | GRAMGRP| NOTE | RE | TRANS | USG |XR
```

Figure 2.1

[3] The TEI is sponsored by the Association for Computers and Humanities (ACD), the Association for Computational Linguistics (ACL) and the Association for Literary and Linguistic Computing (ALLC).

[4] The dictionaries studied are mono and bilingual. The Pequeño Larousse Ilustrado has been used in the case of Spanish.

[5] Basically, typographical markup tags ,editorial and to mark data type: names, numbers, dates, abbreviations, etc.

[6] To simplify and to facilitate the comparison with the definitions of figure 3.2, we have used the notation EBNF, instead of the original notation SGML.

```
<entry>
        <!-- ... information common to both senses -->
        <sense n='1'>
                <!-- ... sense number 1 -->
        </sense>
        <sense n='2'>
                <!-- ... sense number 2 -->
        </sense>
</entry>
```

Figure 2.2

3. Electronic dictionaries: the DRAE

MRDs have been used until now as a source of extracting lexical information because of their varied features which include: (1) they contain a great amount of lexical knowledge about one or various languages; (2) they have a predetermined internal structure: an entry network in which each entry contains structured information (3) they have a certain degree of coding[7] (4) they contain internal relationships between the different lexical objects, both explicit and implicit (5) they have a restricted Lexis; and (6) they follow a fixed methodology in the definition production which enables the identification of the different constituent pieces of the generic term and modifiers.

Furthermore, these features allow another type of processing; the markup to describe and structure the entry information. The explicit information provides a scheme of data for the precise, systematic description of this information.

The markup model must be similar to the organisation model of the MRD lexical information. These dictionaries are based on a hypertext model: a node network with textual information, all connected with links. The hypertext is suitable for materials in which extracting information requires a great amount of cross references. The advantages include quick navigation among information, improved effectiveness and efficiency in searching operations and the possibility of using the information found as an entry for other applications. [Balasubramanian, 1994].

The MRD network nodes store the dictionary entries, following a fairly regular structure. The links represent the relationship among different lexical objects of the dictionary: entries, etymologies, senses, etc., even documents outside the dictionary. The most simple relationship is the equivalence between spelling forms and enables almost all words and multiword of the dictionary to be linked to their corresponding entry. MRDs, therefore, have a poorly structured (or non-existing) macrostructure and a highly structured microstructure.

The electronic format version of the DRAE came out in 1995 and is built from the twenty first edition of the 1992 printed dictionary. The organisation of entry information is not completely regular in spite of the dictionary's more or less fixed scheme. The relationships among the different lexical objects are not explicit, apart from references and optional references. This hinders the automatic extraction of information and is a source of possible inconsistencies which could arise if referential elements are deleted.

The DRAE entry structure is not completely hierarchical[8] since it contains cycles. Figures 3.1 and 3.2 show the scheme of the entries. Figure 3.1 represents graphically the general structural elements of the first level of the hierarchy. Figure 3.2 describes using the notation EBNF, the first two levels of the hierarchy. although the first level shows a hierarchical structure, the scheme contains cycles, as can be seen in the definition of the element FORMA_COMPLEJA[9].

---

[7] Abbreviations, for example, can be considered as predefined codes about parts of speech, information use etc.

[8] Following the hypertext design methodology proposed by Thuring et al [Thuring, 1991], the node structure, in our case, of the entries, should preferably be hierarchical.

[9] FORMA_COMPLEJA is a phrase where the headword is included.

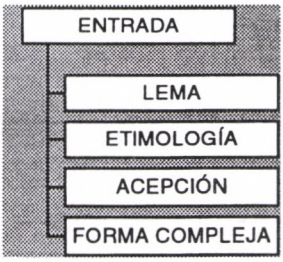


Figure 3.1

```
ENTRADA ::= LEMA (ETIMOLOGÍA)? (ACEPCIÓN)+
            (FORMA_COMPLEJA)*
LEMA::={x | x is a dictionary headword}
ETIMOLOGÍA ::= (unidadLexica | unidadLexicaE | LENGUA |
            ABREVIATURA | ÉTIMO | TRADUCCIÓN | ENCICLO |
            REMISION)*
ACEPCIÓN ::= (INFO_GRAM)* (ESTILO | FRECUENCIA |
            ETIMOLOGÍA | CRONOLOGÍA | GEOGRAFÍA |
            MATERIA | REGISTRO | DEFINICIÓN| ENCICLO |
            COLOCACIÓN | USO | EJEMPLO | REMISIÓN |
            ABREVIATURA) *
FORMA_COMPLEJA ::= SUBLEMA (ACEPCIÓN)*
```

Figure 3.2

There are four types of links in the DRAE: (1) compulsory references, which are a synonymous relationship and optional references which are the combination of the headword with other words; (2) equivalence between spelling forms, which allows jumping from any word to its definition. These links make no distinction between homographs with a different part of speech[10]. (3) Manual jumps. When there is no corresponding definition of a word, the user must intervene and choose from a list of similar spelling forms the spelling he is looking for[11] Finally, in the case of homonyms, the DRAE provides different entries for each of them. In this case (4), the link connects the word to several entries. Here too, the user has to intervene and chose from a contextual menu the entry he wants to display. The user has to intervene to chose the desired entry, which, a priori, is almost impossible, without a thorough knowledge of the dictionary.

Figure 3.3 shows the DRAE entry, **sobornar** (to bribe) in which the structure is described through a markup scheme, partially described in figure 3.2.

```
sobornar.
Del lat. subornare
1. tr. Corromper a alguien con dadivas para conseguir de el una cosa
<ENTRADA>
        <LEMA>sobornar.</LEMA>
        <ETIMOLOGIA>Del <LENGUA>lat.</LENGUA> <ETIMO>subornare</ETIMO></ETIMOLOGIA>
        <ACEPCION>
                <ORDEN>1.</ORDEN>
                <INFO-GRAM><SUBCATEGORIZACION>tr.</SUBCATEGORIZACION>
                <DEFINICION>Corromper a alguien con dadivas para conseguir de el una
                cosa</DEFINICION></ACEPCION>
</ENTRADA>
```

Figure 3.3

4. Application of the SGML/TEI model to the DRAE

The application of the SGML/TEI model to the DRAE has enabled us to incorporate a data scheme which structures and explicitly describes the DRAE entries. The model is applicable since the TEI model tags and the dictionary entry data scheme are sufficiently similar. This similarity can be observed by comparing figure 2.1 with the structural description of the DRAE, shown in figures 3.2 and 3.3. Furthermore, as already explained in point two, the TEI model is considered suitable since among other features, it has sufficient descriptive power. However, this parallelism is not

[10] The conjugated forms of the verb do not refer to the infinitive, which is their related entry. However, those verbal forms coinciding with a dictionary headway have an erroneous link to an entry.

[11] This is not particularly useful, since in many cases the nearest words are not connected at all.

always described in the greatest detail. Due to this, and the lack of systematisation in the information entry presentation, the TEI model has been applied using a methodology based on two criteria. Firstly, the design must be carried out following consecutive levels of refinement. Secondly, simplicity and regularity factors should predominate the degree of scheme descriptiveness. The main advantage of explicit information markup as oppose to the implementation using database approach is the possibility of refining the scheme during the building of the dictionary or lexical database.

In a first phase we have used an initial markup scheme to which successive subschemes have been added. This initial scheme describes the information in a very general way, using only, tags defined as top level elements by the TEI. The following refinement phases depend on the degree of description required, and is in any case limited by the applicability of the coding scheme.

In Table 1 some of the TEI tags comprising the initial scheme applied to DRAE are shown.

Figure 4.1 shows an example of the application of the TEI model to the DRAE.

ELEMENT	*ATTRIBUTES*	*VALUES*	*DRAE*
<superentry>			Groups different homographs[12]
<entry>			Entry
	type		Entry type
		main	Main entry
		hom	Homograph
		xref	For graphic and/or pronunciation variants, but the DRAE pronunciation of the headword is implicit
		affix	For affix entry
		abbr	Acronym
		foreign	Foreignisms (latinisms)
<sense>			Sense
	level		Sense level
<def>			Contains definition text
<eg>			use examples
<etym>			Marks a block of etymological information
<form>			headword (spelling form of the headword)
	type		Headword classification. It is also implicit, thus, it depends on the linguistic knowledge of the codifier
<trans>			Contains the translated text and other type of information related to the translation (morphological, definitions, literal translation, etc)
<xr>			To group all information related to cross references. The reference can be a phrase, sentence or icon, referring to information inside the same document or an external document
<usg>			Information of use

Table 1

[12]In the DRAE, the homographs have different entries with different superindexes, so the tag *<hom>* can be eliminated. However the attribute *type* with value *hom* must be used to describe this type of information. All homographic type entries, will be grouped under the element *superentry*. The TEI, proposes another alternative, to structure them with a single entry, which contains as many elements *hom* as there are homographs. However, this alternative does not correspond to the DRAE structure.

```
<entry>
        <form type='lemma'><orth>sobornar</orth></form>.
        <etym>Del <lang>lat.</lang> <mentioned>comparare</mentioned></etym>
        <sense n='1'>1.
                <gramGrp><pos norm='verbo'></pos><subc norm='transitivo'>tr.</subc></gramGrp>
                <def>Corromper a alguien con dádivas para conseguir de él una cosa</def>
        </sense>
</entry>
```

Figure 4.1

The application of the TEI model in the particular case of DRAE, has moreover enabled us to solve some specific lexicographic problems of this dictionary: (1) a certain irregularity in the presentation of the entry information[13]; (2) irregularities in the design of the dictionary plan[14]; (3) ambiguous definitions or descriptions (such as the etymological description of the word **abedul** (birch). (4) The inaccuracies in the compulsory references to entries with various senses[15] or to different homographs. These lexical problems detected during the entry lexical information markup have been easily solved with TEI.

However, as previously explained, the application of SGML/TEI is not a simple task. Firstly, since it is a wide, overflexible model, it has to be restricted and adjusted to the dictionary to be coded. An in-depth knowledge both of the TEI model and the dictionary is needed. This implies a considerable effort, since in the markup of the dictionaries alone, the TEI has more than 40 basic tags. Moreover, the structure of the model, organised in various intercombinable models[16], must be known. As for the dictionary itself, the DRAE comprises 83,000 entries while our entry description, only at top level, comprises 35 rules.

Secondly, an optimum data scheme is necessary to mark the lexical information correctly and coherently. This scheme will not only facilitate the automatic information markup but will also allow data coherency to be controlled. An optimum quality data scheme should therefore be: (1) maintainable, that is, capable of tolerating modifications without losing consistency; and (2) complete, that is, capable of representing all the information contained in the entries, both implicit and explicit.

The implicit information, is that which the user infers, based on his own knowledge of the language and of the dictionary[17]in use. The data scheme should explicitly model this type of information. Thus, the most complex implicit information to model is the relationship among elements of the dictionary.

The relationships among lexical objects of the dictionary can appear marked or not. The marked ones stand out either typographically or using abbreviations.[18] They are basically the references inside the senses to the headword they define. The references are links to information under another entry, either compulsory, since it is necessary to understand the object described, or optional, since it adds more complementary information about the headword. In turn, the references can be links to the headword and to phrases that include the headword[19]. In the electronic version of the DRAE, the references to FORMAS_COMPLEJAS (phrases) are not always satisfactorily solved. The coding of

---

[13] The etymology of a word is not explicitly indicated, it is part of the definition. Example **leopoldina, federica (a la)**

[14] For example, the etymological information could appear dispersed in different elements of the entry. In the entries of some derivatives of the word ***casa, (caserón, casilla)*** it can be seen that the etymological information has not always been allocated its usual place, before the senses.. In the entry ***caserón*** the etymology appears inside the sense.

[15] For example, in the word ***desmentido, da***, in the third sense the synonymous definition ***mentís*** is given, without specifying to which sense of ***mentís*** it refers.

[16] A main module, the basic modules and optional modules.

[17] As described in its Preliminares (preliminaries).

[18] As shown in the second sense of the headword ***ad nútum*** in figure 4.2, in many cases they are preceded by marks such as V. (see) In this case the typographic mark "semibold" is used.

[19] these phrases are named FORMAS_COMPLEJAS in our DRAE grammar.

these relationships with the TEI model is direct and easily solves this type of irregularities. In figure 4.2 a coding example of this type of relationships: references to a FORMA_COMPLEJA (phrase)[20].

ad nútum.
1. expr. lat. a voluntad.
2. V. **beneficio amovible ad nútum.**

```
<sense n='2'>2.
    <xr type='sublema'>
        <lbl norm='Véase'>V.</lbl>
        <ptr target='beneficio.amovible ad nútum'>beneficio amovible ad nútum</ptr>
    </xr>
</sense>
```

Figure 4.2

The unmarked relationships are more difficult to define and classify because of their varied nature and are left up to the user's linguistic competence. The TEI model can solve those relationships concerning the spelling[21]. The information about variants in the DRAE, for instance, can only be found in forms favoured by the Academia and are therefore sent to the favourite and defined form[22]. The TEI model, however, does not foresee another type of relationship, principally, of a semantic nature.[23] This lack is important both, if the dictionary is used as a source of lexical knowledge to build computational lexicons, and if it is used as a tool for linguistic research.

5. Remarks and Future Work

In this article, we have described the work we are currently carrying out within the framework of our research group, whose general aim is the study and development of text comprehension systems. This research is directed towards the building of Spanish MRDs, which can include lexical information in a semi-automatic form. This information is obtained form existing MRDs and in particular the DRAE. The approach we are following is based on explicitly defining the structure and all (implicit and explicit) relationships contained in the lexical information of a MRD, through a markup language. Until now our most important results can be resumed as follows: (1) The explicit markup of the entry lexical information, provides a systematic, precise description of that information. (2) Other advantages of using a model based on standard SGML form markup include independence of the markup information of hardware and software. For these reasons, the automatic extraction of lexical information can be greatly improved. (3) The application of a data schema implies the standardisation of this information and it entails the detection and correction of inconsistencies in the dictionary. Therefore, it can solve some lexicographic problems. However, (4) the TEI model must be enlarged to include certain implicit information, such as semantic relationships among dictionary elements. Furthermore, the TEI model must be restricted using a data scheme which is suitable form the representation of lexical information. Finally, (5) The insertion of a data scheme through a descriptive markup language, allows and simplifies the modifications on this scheme during the building of the dictionary. These modifications imply, only, a tag change in a text file.

Currently, we are completing the DRAE entry coding scheme in two directions. Firstly, by including the semantic relationships, already stated. For this purpose, we use other Spanish

[20] It should be emphasised that although it is not shown, the target coding of that reference (the complex form ***amovible ad nútum)*** is codified with a tag <re> which should have a 'sublema' value in its attribute type. In this way, the attribute type values of the reference tag <xr>and the target tag <re> are consistent.

[21] The TEI proposes only 9 tag types for the relationships among objects .These tag attributes only describe the variations among spelling and phonetic forms of the referential and referred to objects.

[22] The entry ***sicología*** refers to ***psicología*** (preferred form), however in ***psicología*** there is no information about the existence of the variant ***sicología.***

[23] Basically synonym, hipernym/hiponym, holonym/meronym, antonym

language dictionaries, where the semantic relationships are more explicit. Secondly, by refining the DRAE entry structure, to obtain the most detailed description possible, without forfeiting simplicity in the data scheme, and consequently, the applicability of this scheme. Furthermore, we are developing an application form automatic entry encoding, using the basic scheme already stated.

6. References

AQUILEX. (1995). Reusable Multilingual Lexicons for Natural Language Processing. *AQUILEX-EC-US004*. [http://www.fcul.research.ec.org/esp-syn/text/ec-us004.html]

Balasubramanian, V. (1994). *State of the Art Review on Hypermedia Issues and Applications.*[http://cbl.leeds.ac.uk/nikos/tmp/hypemedia/hypemedia.html]

Boguraev, B. and Pustejovsky, J. (Eds.) (1996). *Corpus Processing for Lexical Acquisition*. The MIT Press.

Calzolari, N. (1994). European efforts towards standardising language resources. In P. Steffens, editor. *Machine Translation and the Lexicon*. Springer-Verlag.

Castellón, I. (1992). *Adquisición Automática de Conocimiento Léxico*. Tésis Doctoral. Departamento de Filología Románica. Barcelona.

DRAE (1992). *Diccionario de la lengua española*. Real Academia Española. Madrid.

EAGLES. (1996). Computational Lexicons Working Group. Reading Guide. *EAGLES Document EAG-CLWG-FR-2*. [http://www.ilc.pi.cnr.it/calz/calz.html]

Elmasri, R., Navathe, S.B. (1997). *Sistemas de Bases de Datos. Conceptos Fundamentales*, (2ª edition). Addison-Wesley Iberoamericana.

Goldfarb C.F., (1990). *The SGML Handbook*, Oxford University Press.

Gonzalo, J.; Verdejo, M.F.; Chugur, I.; López F.; Peñas, A. (1998). Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet. *Actas del XIV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'98)*, Alicante, págs. 97-103.

Grishman, R; Macleod, C; Meyers, A. (1994). Comlex syntax: building a computational lexicon. In *Proceedings of the 15th Annual Meeting of the Association for Computational Linguistics, (Coling'94)*. 268-272. Kyoto. Japan.

Grover, C; Carroll, J; Reckers, J. (1993). The Alvey Natural Language Tools grammar (4th realese). *Tecnical Report 284. Computer Laboratory, Cambridge University*, UK.

Guthrie, L, Pustejovsky, J; Wilks, Y; Slator, B.M. (1996). The Role of Lexicons in Natural Language Processing *Communications of the ACM*. Vol 39, No 1

Ide, N.M. and Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities 29* (2).

Kaplan, R.M. and Bresnan, J. (ed.).1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Bresnam J. (ed.) *The Mental Representation of Grammatical Relations*. Cambridge, Mass: MIT Press.

Martí, M.A.; Castellón, I.; Fernández, A. (1998). Extracción de información de corpus diccionariales. *Novática*, Mayo-Junio

Miller,G.A.;Beckwith,R; Fellbaum,C.;Gross;D.;Miller,K; Tengi, R. (1993). Five papers on WordNet. *CSL Report 43*, Princenton University, Cognitive Science Laboratory.

Sperberg-McQueen, C.M. (1995). The TEI: History, Goals and Future. *Computers and the Humanities* 29, 5-15.

Sperberg-McQueen, and C.M.; Burnard, L. (1994) *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Chicago and Oxford: Text Encoding Initiative, 1994. [ftp://ftp.tei.uic.edu/pub/tei]

Thuring, M.; Haake, J.; Hannemann, J. (1991). *Hypertext'91 Proceedings. ACM* Press

Varile,N.; Zampolli, A. (Ed.). (1992). *COLING92 International Project Day*. Giordini Editori, Pisa.

Yokoi, T. (1995). The EDR Electronic Dictionary. *Communications of the ACM*. Vol. 38, No. 11.

Semantic Tagging: A Survey

THIERRY FONTENELLE

Abstract

Word Sense Disambiguation (WSD) is probably one the most complex tasks developers have to solve when designing NLP applications. It is not just an engineering problem since, apart from choosing (or constructing) the right type of lexical resource, it entails a number of almost philosophical choices which depend heavily on the approach to the basic problem of word meaning. This paper will address some of the problems related to semantic tagging and will describe some of the resources built for particular applications. We will argue that the type of resources needed depends on the nature of the application the developer has in mind. Our aim is not to provide an exhaustive list of projects and resources, but rather to select a few such resources in order to bring out the various facets of the problem and sketch possible solutions.

1. WORD SENSE DISAMBIGUATION: PHILOSOPHY OR ENGINEERING?

Philosophers ranging from Aristotle to Wittgenstein have addressed the question of meaning (and in particular the meaning of words) for centuries. It is not our ambition to contribute to the advancement of these philosophical debates. But it is not possible to ignore their existence either, and even if one is not inclined to embark on high-level, abstruse discussions, it must be admitted that some 'hot topics' in linguistics, and more particularly in lexical semantics, owe a lot to the contributions that were made several centuries ago. The theory of the so-called 'qualia structures', which is currently gaining ground in the framework of Pustejovsky's Generative Lexicon (Pustejovsky 1995), is a case in point: it claims to provide the researcher with a novel type of semantic analysis which can supposedly cope with the creation of new senses from a finite number of word senses. This theory can be traced back to Aristotle's *qualia*, which Pustejovsky rediscovered as a possible way of solving some of the most vexing problems in computational linguistics.

Recent advances in natural language processing enable us to break down linguistic processing into a number of tasks ordered as a function of their sequential use (see Grefenstette 1996, 1998 for more details):

(1) tokenising/lemmatising (morphological analysis)

(2) part-of-speech tagging

(3) parsing (syntactic tagging)

(4) word sense tagging (semantic tagging)

The first three tasks use relatively well-known techniques and it can be claimed that some consensus has been reached as to how the performance of a system designed to carry them out can be evaluated. This is especially true of morphology and of part-of-speech tagging. At the other end of the scale, however, word sense tagging is still in its infancy. One of the reasons is that the appropriate linguistic resources for large-scale semantic tagging are not widely available. Another hurdle is the difficulty of agreeing on whether the answer produced by a WSD system is correct or not. As noted by Kilgarriff (1997), anyone who has ever compared the definitions and meaning distinctions in different dictionaries for a given word cannot have failed to notice the lack of agreement when it comes to breaking down a word into senses. Kilgarriff even goes as far as to argue that there is no such thing as word senses. His contention is that the basic unit for WSD cannot be 'word senses', because the notion is not sufficiently well defined. In contrast, the basic units are occurrences of a word in context and these corpus citations are to be grouped into one or several word clusters forming a distinct word sense. Following this approach, a crucial implication is that the performance of a WSD system depends on the task to which it is put. Kilgarriff convincingly demonstrates that "word senses are undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not worth making" (1997).

This approach to word senses seen as abstractions over clusters of word usages contrasts with the widely held view that words have listable, enumerable meanings. The latter view is shared by most of the researchers who use the WordNet lexical database in their experiments (Fellbaum 1998). The scarcity of lexical resources for WSD probably explains why WordNet is the most widely used dictionary, even if, in addition to a number of errors and inconsistencies, the sense distinctions it makes may sometimes be too fine-grained for some applications.

2. SEMCOR: A SENSE-TAGGED CORPUS

The WordNet lexical database, which is publicly available, was used to build semantic concordances from two textual corpora. Landes *et al.* (1998) show how these corpora were annotated with information on which WordNet sense is used for each word. These semantic concordances were constructed manually by a team of 'taggers' and a specific software package was designed to facilitate the coders' task. Of course, when the target word had only one sense in WordNet, the job was fairly easy since it entailed checking that the WordNet sense was the appropriate meaning in the context of the sentence under scrutiny. More interesting were the polysemous words which account for 83% of the open-class words in the Brown corpus. A typical sentence with semantic tag information looks as follows:

Source sentence:

- From what I was able to gauge in a swift, greedy glance, the figure inside the coral-colored boucle dress was stupefying.

Tagged sentence:

<p>
<s snum=87>
<wf cmd=ignore pos=IN>From</wf>

<wf cmd=ignore pos=WP>what</wf>
<wf cmd=ignore pos=PRP>I</wf>
<wf cmd=done pos=VB lemma=be wnsn=1 lexsn=2:42:03::>was</wf>
<wf cmd=done pos=JJ lemma=able wnsn=1 lexsn=3:00:00::>able</wf>
<wf cmd=done pos=VB ot=notag>to</wf>
<wf cmd=done pos=VB lemma=gauge wnsn=1 lexsn=2:31:00::>gauge</wf>
<wf cmd=ignore pos=IN>in</wf>
<wf cmd=ignore pos=DT>a</wf>
<wf cmd=done pos=JJ lemma=swift wnsn=1 lexsn=5:00:00:fast:01>swift</wf>
<punc>,</punc>
<wf cmd=done pos=JJ lemma=greedy wnsn=1 lexsn=5:00:00:acquisitive:00>greedy</wf>
<wf cmd=done pos=NN lemma=glance wnsn=1 lexsn=1:04:00::>glance</wf>
<punc>,</punc>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=figure wnsn=2 lexsn=1:08:00::>figure</wf>
<wf cmd=done pos=IN ot=notag>inside</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done pos=NN lemma=coral wnsn=1 lexsn=1:07:00::>coral</wf>
<wf cmd=done pos=JJ lemma=colored wnsn=1 lexsn=3:00:00::>colored</wf>
<wf cmd=done pos=NN lemma=boucle wnsn=1 lexsn=1:06:00::>boucle</wf>
<wf cmd=done pos=NN lemma=dress wnsn=1 lexsn=1:06:00::>dress</wf>
<wf cmd=done pos=VB lemma=be wnsn=1 lexsn=2:42:03::>was</wf>
<wf cmd=done pos=JJ lemma=stupefying wnsn=1
lexsn=5:00:01:impressive:00>stupefying</wf>
<punc>.</punc>
</s>

As can be seen above, the format follows SGML guidelines with sentence <s> elements nested within paragraph elements. Information about word form, syntactic category and semantic tags is represented in terms of attribute-value pairs (lemma referring to the base form of a word, as in *was* ← *be* above; lexsn representing the location of the sense in the database and wsn specifying the word sense number). In the example above, the adjective *greedy* corresponds to the 1st word sense in the lexical database. The full entry with its sense distinction is given below:

WordNet entry for adjective greedy:

3 senses of greedy

Sense 1

avaricious, covetous, grabby, grasping, greedy, prehensile -- (immoderately desirous of acquiring e.g. wealth; "they are avaricious and will do anything for money"; "casting covetous eyes on his neighbor's fields"; "a grasping old miser"; "grasping commercialism"; "greedy for money and power"; "grew richer and greedier"; "prehensile employers stingy with raises for their employees")

=> acquisitive (vs. unacquisitive) -- (eager to acquire and possess things especially material possessions or ideas; "an acquisitive mind"; "an aquisitive society in which the craving for material things seems never satisfied")

Sense 2

avid, devouring (prenominal), esurient, greedy -- ((often followed by `for') ardently or excessively desirous; "avid for adventure"; "an avid ambition to succeed"; "fierce devouring affection"; "the esurient eyes of an avid curiosity"; "greedy for fame")
=> desirous (vs. undesirous) -- (having or expressing desire for something; "desirous of high office"; "desirous of finding a quick solution to the problem")

Sense 3

greedy -- (wanting to eat or drink more than one can reasonably consume; "don't be greedy with the cookies")
=> gluttonous (vs. abstemious) -- (given to excess in consumption of especially food or drink; "over-fed women and their gluttonous husbands"; "a gluttonous debauch"; "a gluttonous appetite for food and praise and pleasure")

Several remarks are in order here. First of all, it should be borne in mind that the whole tagging process is entirely manual, which means that the quality of the resulting corpus is heavily dependent on the lexicographical skills of the people who tagged the corpus (the taggers). Obviously, as Landes *et al.* (1998:212) point out, such a concordance can be viewed as a corpus of syntactically and semantically disambiguated text. The designers also argue that it can be used as a lexicon where many of the word senses are illustrated by example sentences. Whether it can equally be employed to acquire information on the relative frequency of word senses in written text is a moot point, however, since the size of the corpus is limited (the Brown corpus comprises 1 million words only and is fairly old), and the technique is not automatically applicable to other corpora without a huge amount of manual work. Moreover, the identification of multi-word units by the tokeniser is limited to contiguous constituents. Phrasal verbs as in *He rang me up* have to be identified manually by the tagger before the appropriate WordNet sense (for *ring up*, distinct from the senses of *ring*) can be assigned. On the other hand, Leacock *et al.* (1998) suggest a method for re-using WordNet relations to identify the correct sense of a word in context, which means that the potential of the database has yet to be fully exploited. In their experiment, Leacock and her colleagues use a special program to retrieve from WordNet all the monosemous relatives of a polysemous word sense. The program then retrieves a collection of example sentences which contain these monosemous relatives and 'daughter' collocations containing the target word as context (e.g. for the polysemous noun *court*, they extract example sentences containing occurrences of *tribunal*, of the monosemous synonyms of *court* listed in WordNet or of daughter collocations such as *superior court* (1998:160-163)). They are aware that manually tagged training data are not available in unlimited quantities and their approach enables them to collect training examples from bigger corpora by exploiting the spectrum of WordNet relations. The statistical comparison of the contexts of the monosemous relatives and of the polysemous target words makes it possible to automatically assign a word sense to the target word.

3. FRAMENET

The Berkeley FrameNet project is another attempt at describing word senses using corpus evidence (Baker *et al.* 1998). The originality of the lexical database it aims at creating is that the description of each word includes, in addition to 'traditional' morpho-syntactic information, a list of all possible constellations of frame elements. Moreover, each word sense is linked to a set of corpus-derived sentences that have been annotated with frame-semantic information, which is a form of linguistically elaborated sense tagging (Fillmore & Atkins 1998). For want of space, I cannot go into the details of the underlying theory of frame semantics here (see Fillmore & Atkins 1994). Suffice it to say at this stage that frames are general or specific

conceptual structures involving entities (frame elements) which participate in these structures. The aim of a frame-based lexicon is to describe the combinatory potential of a lexical item, which amounts to specifying how each frame element (another name for what some linguists would call case roles or theta roles) is realised at the surface level. Examples of frames are, for instance, the DRIVING frame, which involves the following elements: a DRIVER (= a primary MOVER), a VEHICLE, optionally secondary movers such as CARGO or RIDERS. If one starts from the hypothesis that the DRIVER or the VEHICLE can appear as subjects, that VEHICLE, CARGO and RIDER can be realised as direct objects and that the PATH and VEHICLE elements can surface as oblique complements, we have the necessary apparatus at our disposal to account for the following sentences (see Baker *et al.* 1998):

- I[DRIVER] drove my friends[RIDER] home[PATH]
- The truck[VEHICLE] drove down to the Riviera[PATH]
- John[DRIVER] used to drive his BMW[VEHICLE] at lightning speed around the city[PATH]
- They[DRIVER] were cycling along at breakneck speed.

Another frame, activated in a different semantic field altogether, is the so-called HEALTH frame (Lowe *et al.* 1997). The typical frame elements which can be identified in a health care scenario are the following ones:

Frame element	Meaning
Healer	Individual who tries to improve the health of a patient
Patient	Individual affected by a disease
Disease	Sickness which should be removed
Wound	Tissue damage to the body of the patient
Body Part	Limb or organ affected by the disease or wound
Symptom	Evidence showing the presence of a disease
Treatment	Process aiming at causing recovery
Medicine	Substance used to bring about recovery

The following sentences show that the subject of verbs such as *recover*, *cure*, or *heal*, which participate in the health frame, can correspond to a variety of frame elements:

- The doctors were not able to *cure* his cancer. [S=HEALER]
- A couple of aspirin *cured* my headache. [S=MEDICINE]
- Her ankle *healed*. [S=BODYPART]
- The scratch *healed* rapidly. [S=WOUND]

The aim of the FrameNet project is to tag the items of a given frame in terms of Frame Element Groups, i.e. lists of frame elements occurring in a phrase or in a sentence headed by a given word (Lowe *et al.* 1997:5). In a sentence such as:

- The dermatologist *treated* my skin with ultraviolet radiation.

The verb *treat* occurs with the following Frame Element Group:

{H,B,T} Healer (= dermatologist); Body Part (= skin); Treatment (= ultraviolet radiation)

Special software was designed to enable the lexicologists to tag each sentence of the corpus with frame semantic annotations. Exploiting the potential of graphical and colour interfaces, the corpus query tools make it possible to visualise the structure of concordances in which different colours are used to highlight different frame elements (e.g. all healers would appear in blue while body parts would appear in red). Once the concordances have been examined and the tagging process is finished, the analyst is able to formulate sophisticated queries on the corpus. A typical request would be, for instance:

Retrieve all occurrences of the verb *cure* featuring the Frame Element Group {H,D} (= Healer, Disease), which could yield the following sentence:

- The doctor *cured* his arthritis.

In contrast, a query on occurrences of the verb *cure* with {M,W} (Medicine, Wound) would extract the following sentence (from Cobuild 1987):

- It was used as a folk medicine to *cure* snake-bite.

Interestingly, as Lowe *et al.* 1997 note, this approach makes it possible to investigate some linguistic phenomena such as ergativity (a.k.a. the causative/inchoative alternation) from a different angle. Corpus evidence indeed shows that the verb *heal* can be used both causatively (transitively) and inchoatively (intransitively) only when it involves a wound as entity that changes state (*the ointment healed the cut* vs. *the cut healed*). A transitive use involving a disease (as in *the sorcerers healed my arthritis*) has no inchoative/intransitive counterpart (* *my arthritis healed*).

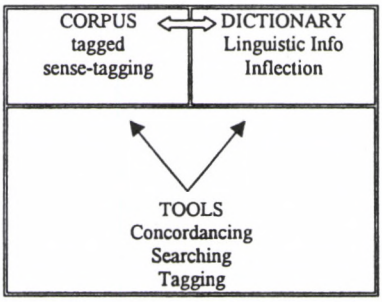
A major drawback of the interesting approach advocated by FrameNet is that it relies on the manual interpretation of corpus material by highly skilled linguists and lexicographers. The resulting hand-tagged corpus examples are obviously of invaluable help in a word sense disambiguation perspective and the theoretical framework enables an extremely detailed and fine-grained analysis of the combinatory potential of lexical items in a given semantic area. It is hoped that the FrameNet proponents will manage to make good use of other types of lexical resources as well in order to at least partially automate the complex tag-assignment tasks. It should indeed be possible to use some of the categories of thesauri such as Roget's or WordNet in order to derive lists of lexical items belonging to some categories of frame elements (for the word *doctor*, for instance, WordNet proposes hyponyms and related terms such as *dermatologist, surgeon, cardiologist, orthopaedist, gynaecologist*, ... which can all be used as exponents of the 'healer' frame element in the health frame).

4. CIDE

A different approach to WSD consists in re-using a commercial dictionary to tag corpus data. Instead of embarking on the construction of a special hand-crafted lexicon with an army of pseudo-lexicographers (generally a bunch of underpaid undergraduate students, as in too many commercial projects), many researchers have tried to automate the lexical acquisition problem by tapping data contained in the machine-readable versions of commercial dictionaries (and

especially English learners' dictionaries). This idea is not new and can be traced back to the early 1980s with Amsler's seminal work on the Merriam-Webster dictionary (Amsler 1980) or Michiels's exploitation of the LDOCE database (Michiels 1982; see also Wilks *et al.* 1996 for a useful history of MRD-based research). Ide & Véronis (1998:9-11) sketch a number or problems related to the automatic extraction of large knowledge bases from MRDs: they show that this initial goal was not fully achieved (the extraction of taxonomies from dictionary definitions, for instance, has proved to be fraught with insuperable problems). The numerous inconsistencies which mar these dictionaries are now well known and very few linguists today would argue that MRDs are the sole source of lexical data usable for WSD. Since the advent of large computer corpora over a decade ago, many linguists have now turned to these large bodies of evidence in their quest for lexical knowledge. The argument held by corpus linguists is that only corpora can provide us with a correct picture of the way words behave and combine with other words. The sheer amount of concordances to be examined by a lexicographer has created a situation in which "the lexicographer is like a person standing underneath Niagara Falls holding a rainwater gauge, while the evidence sweeps by in immeasurable torrents", to quote Church *et al.* (1994:153). This profusion of data accounts for the development of tools and techniques to sift through corpus data, to arrange and sort KWIC lines, to focus on statistically significant phenomena (by resorting to well-known techniques such as MI calculations, for instance - see, inter alia, Atkins 1982, Clear 1994...). One immediately sees the snag here: in order to get the most out of corpus data, the researchers are forced to process the corpus with tools ranging from simple tokenisers to part-of-speech taggers to robust parsers. When it comes to performing semantic analyses, some kind of word sense disambiguation is most often required and is equally often hampered by the lack of appropriate resources. The researcher then finds himself in a sort of Catch-22 situation since the construction of a computational lexicon based on corpus data requires the use of techniques which draw on the availability of large-scale lexical resources which are themselves one of the ultimate goals of the exercise. Of course, one can only agree with Grefenstette (1998) who argues that a lot can be achieved with approximate linguistic tools. But Grefenstette is mainly concerned with abstracting away from surface differences in text and his approach probably works best when it is limited to the structural processing of texts (morphological analyses, POS tagging, shallow parsing...). The highest level of abstraction, that of semantic tagging, is still in its infancy because, as he puts it, "the problem is no longer one of simple structure, but also of meaning" (1998:36). It is therefore interesting to see how some researchers have tried to tackle the vicious circle alluded to above by using a dictionary produced on the basis of a big corpus to tag the corpus itself and hence refine the dictionary in a spiral-wise fashion.

The CIDE dictionary (Procter 1995) is the result of such an approach where a corpus was used to compile a dictionary and the dictionary itself was used to enhance the performance of the corpus tools. This interactive way of using a corpus and a dictionary which mutually enrich each other can be illustrated as follows:



CIDE is a learner's dictionary whose definitions are written with a controlled defining vocabulary. Each word sense is assigned a number of grammar codes which reflect the environment into which the lexical item can be inserted. Semantic features for nouns and selection restrictions (for verbs) have been added, although they do not appear in the printed version of the dictionary. These and other features, such as subject field codes, are available in the CIDE+ electronic database which can be obtained from Cambridge University Press (http://www.cup.cam.ac.uk/elt/reference/data.htm). An interesting innovation, which proves to be quite useful in a WSD perspective, is the use of a Guideword, i.e. a general indicator which can appear at sense level to guide the user to the relevant entry. As can be seen below, the guideword is a little more than just a hyperonym. It can in fact be viewed as a combination of superordinate and subject field. Consider the entry for *port*:

port TOWN *n* a town by the sea or by a river which has a harbour ... or the harbour itself

port CONNECTION *n* [C] *specialized* a part of a computer where wires can be connected in order to connect other pieces of equipment, such as a printer

port LEFT *n* [U] *specialized* the left side of a ship or aircraft when viewed from a position inside when you are facing the front

port WINE *n* [U] a strong sweet typically dark red wine made in Portugal

port BAG *n* [C] a case or bag

What makes CIDE interesting is that the guideword is accessible without requiring any linguistic processing of definition texts to extract genus words. In the electronic version of the dictionary, this guideword is surrounded by SGML tags which make it readily identifiable. SGML tags are also used to house information about subject fields and general domain indicators. Consider the following entry corresponding to the fourth sense of *port*:

<sense><headword><record><key>port* 4* 0</key>
<word-group><word>port</word><pos>n</pos><grammar>U</grammar></word-group>
<guideword>WINE</guideword><subj>VIN</subj><class>Drink</class>
<def>a strong sweet typically dark red wine made in Portugal</def>
</record></headword></sense>

The third line in the excerpt above is especially interesting since it contains indications about the guideword (wine), the subject field (*port* is a term used in viticulture - <subj>VIN</subj>) and the general class to which the headword belongs (beverages, since *port* is tagged as <class>Drink</class>). One clearly sees the hierarchy adopted by the lexicographers here and information about the general semantic classes to which a given item in a given sense belongs is undoubtedly extremely useful in a WSD perspective. It makes it possible to specify that *port*_4 may be found in an environment where drinks are expected (e.g. verbs such as *gulp*, *drink*, *swallow* subcategorize for direct objects which should be marked as [+drink], which is true of *port* and other wines, but also of soft drinks, water, whisky, etc.). At the same time, this information is distinct from the guideword, which in this case corresponds to a true hyperonym, or from the subject field code, which can be used for domain identification.

Harley & Glennon (1997) describe a sense tagger which tries to link every word in a large corpus to its corresponding entry in CIDE. Interestingly, the tagging process is not a yes-no assignment since the tagger assigns scores to all possible senses of each word in a sentence. Scores are then reaped and adjusted, for instance if a collocation is recognised (in which case the score is increased). The semantic tag which is assigned by the automatic tagger corresponds to the sense which achieved the highest score. The tagger takes the following information into account:

- information on multi-word units in the CIDE database
- subject domain (the codes assigned to each word in a sentence are collected, compared and noted)
- part of speech
- selection restrictions

Harley & Glennon note that selectional preference pattern matching has proved one of the most useful of all tests. They give the following example: in the sentence *The head asked the pupil a question*, *head* could be a body part, state, object, human or device. *Pupil* could be a human or body part. *Question* could be communication or abstract. But *asked* with two objects can only have the pattern "human asked human communication", which means that all the senses can be correctly assigned by just using the selectional preferences (1997:2).

5. A LEXICAL-SEMANTIC DATABASE

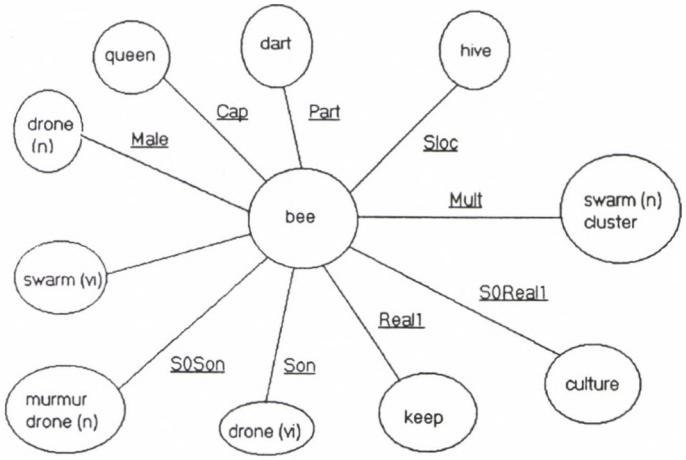
Fontenelle (1997a,b) describes a lexical-semantic database constructed from the machine-readable version of the Collins-Robert English-French dictionary (Atkins & Duval 1978, first edition). A careful analysis of the micro-structure of this dictionary revealed that it contains a treasure trove of collocational (combinatory) information. The database created from the typesetting tape of the dictionary allows diversified access and an original feature of the project is that the dictionary has been enriched with information about the lexical-semantic relationships which link the headwords and the "metalinguistic" markers or indicators appearing at word sense level. In order to cater for the variety of semantic relationships discovered in the dictionary entries, Mel'čuk's system of lexical functions was chosen because it offered a broad spectrum of links uniting a base and its collocates (Mel'čuk 1984). The following examples, excerpted from the printed version of the dictionary, show what material was used to construct the lexical-semantic database:

dart 1 *n* ... **c** *(weapon)* trait, javelot; *(liter) [serpent, bee]* dard
drone 1 *n* **a** *(bee)* abeille mâle, faux bourdon ... **b** *(sound) [bees]* bourdonnement; *[engine, aircraft]* ronronnement, *(louder)* vrombissement ... **2** *vi [bee]* bourdonner; *[engine, aircraft]* ronronner, *(louder)* vrombir...
keep *vt* **e** *(own; look after) shop, hotel, restaurant* tenir, avoir; *house, servant, dog, car* avoir; *(Agr) cattle, pigs, bees, chickens* élever, faire l'élevage de

A quick glance at the three entries above reveals that the word *bee* appears 5 times as a metalinguistic indicator in italics, i.e. as a piece of information provided by the lexicographer to guide the user to the appropriate meaning and the correct translation. The status of this metalinguistic label differs from one entry to the other, however, and the typographic conventions contribute to clarifying this status. *Bee* between parentheses undoubtedly refers to a synonym or a hypernym, s.v. *drone* (1), for instance. When surrounded by square brackets, *bee* can play the part of a collocate or noun complement (drone of bees; dart of a bee) or can refer to a typical subject in a verb entry (bees typically drone). The occurrence of the unbracketed string *bee* in a verbal entry points to a verb-object relation. In addition to these syntagmatic relations (Noun-Noun or Noun-Verb or Verb-Noun collocations), the database includes an explicitation of the lexical-semantic link between a given italicised indicator and the entries under which it appears. In terms of Mel'čuk's lexical functions, the entries containing *bee* above can be rewritten as follows:

Male (bee) = drone (n) (a male bee is a drone)
Son (bee) = drone (vi) (typical verb for the sound made by bees)
S_0Son (bee) = drone (n) (typical noun for the sound made by bees)
Part (bee) = dart (part-whole relationship)
$Real_1$ (bee) = keep (verb denoting the typical action associated with bees)

The semantic network which can be built for all the occurrences of *bee* in the English-French part of the dictionary can be represented diagrammatically as follows:



[Lexical-semantic relations: Cap = head of; Mult = group of; S_{loc} = noun for the typical location of; Son = sound of (verb) ; S_0Son = sound of (noun); $Real_1$ = typical action (verb); S_0Real_1 = noun for the typical action]

The multi-access database can be queried from various angles and it is now possible to retrieve, say, the verbs expressing the typical sound of bees or the nouns denoting a regular group of bees (see Fontenelle 1997a for more details on the structure of the database, the query interface and the potential use of the dictionary).

The interest of such a database in a WSD perspective should be obvious. The collocational link which holds between the verb *keep* and the object collocate *bee* is a crucial item of information: when used in combination with *bee*, *keep* should be translated as *élever* and the prototypical translation (the default sense, i.e. the only translation one finds in pocket dictionaries), *garder*, is not possible in this context. One clearly sees that this syntagmatic relationship ought to be kept track of when parsing the sentence. Admittedly, the nature of the lexical-semantic relationship need not be made explicit and a WSD system using such a database could come up with the correct translation without labelling this relationship in terms of lexical functions *à la* Mel'čuk. There are cases where the explicit labelling of these relations does contribute to disambiguation, however, as is shown in the following sentences:

(1) The teacher drew a map of Sweden.
(2) They shot clouds of arrows at us.
(3) I couldn't install this piece of software.

The *of*-phrases have a different status in (1) and (2)/(3). In the "N1 of N2" structure above, N1 is the semantic head of the phrase in (1) only. In (2) and (3), N2 (*arrow, software*) is the semantic head of the noun phrase and *cloud/piece* act as quantifiers, which in terms of lexical functions, can be represented as Sing (single unit of) or Mult (group of). It is important to capture this link, as shown in Fontenelle (1998), because, in a WSD perspective, the meaning of *draw* in (1) depends on its co-occurring with *map*, and not with *Sweden* (draw + map → dessiner), while *shoot* has the meaning it has in (2) because it takes *arrow* as direct object and, similarly, *install* collocates with *software*, not with *piece*.

The Collins-Robert database was used in the framework of the DECIDE project (Grefenstette *et al.* 1996) and experiments were conducted in order to show how this bilingual dictionary enriched with semantic information could be exploited to retrieve semantically related sentences from a corpus. In particular, Fontenelle (1997b:293-296) shows how the apparatus of lexical functions was exploited to extract semantically coherent sets of concordances from a tagged corpus. The retrieval functionalities of the query interface indeed enable the user to extract pairs of items linked by a particular lexical function (e.g. collocations combining a sound verb and its prototypical subject - *bells ring, bees drone, cocks crow, asses bray* - are linked by the lexical function Son). The resulting list, which comprises hundreds of collocate pairs, was then submitted to a Corpus Query Processor developed at the University of Stuttgart. A flexible query generator using macros and template files for complex search patterns (Schulze 1996) was used to search a subset of the Brown corpus for a noun and a verb separated by an unspecified number of words within sentence boundaries. The noun and the verb in question were of course taken from an input file consisting of data extracted automatically from the Collins-Robert database. The following concordances, which all express the production of a sound, were extracted from the 255,000-word subset of the corpus:

166514: not touching anything . The <clock on the mantel_piece was scandalized and ticked> so loudly that he glanced at it over his shoulder

165923: _from the hall . Between the <telephone and the wall_plug there was sixty feet of cord , and when the conversation came_to_an_end , Eugene carried the instrument with him the whole length of the apartment , to his bathroom , where it rang> three more times while he was shaving and

179779: active imagination . When the <telephone rang> on the day after Hino went down to the village

148882: as shouting triumphantly . A <train hooted> . Instantly , he chilled . They were pursuing him

In the semantic tagging perspective described in this paper, it is easy to imagine how this database could be used to enhance dictionary-based corpus analysis. Instead of starting, as above, from a given lexical function without knowing which words exemplify it, it is possible to start from a particular word and to specify that one is interested in retrieving only concordances which contain this very word used in a particular semantic environment. For instance, as is shown in Fontenelle (1997b:296), a lexicographer who wishes to retrieve sentences referring to a group of trees may not know how this collocational relationship can be expressed. Knowing that the group_of relation corresponds to the Mult lexical function is enough, however, and the semantic network built around *tree* can be searched for the exponents of this function to yield the following list: *clump, cluster, line, row, stand* [of trees]. In the experiments carried out with the Stuttgart MacroProcessor linked to the Collins-Robert database, the following concordances were then obtained:

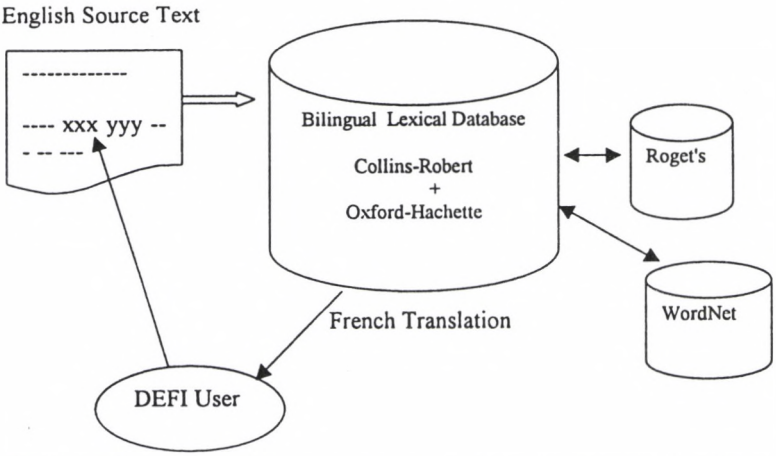
157953: ked with sweat . There was a <clump of trees> that appeared to provide cover right up

215343: st out_of_sight in the first <stand of trees> , fed by a half dozen springs that popped_out of

The application one can envisage is a situation in which a dictionary compiler could be provided with a collection of semantically related instances of a keyword or of a general meaning. In the perspective described above, a lexicographer can search for the occurrences of the word *ring* when it denotes a verb of sound simply because the relevant lexical function (Son) has been used as a primary access key in the database. Such very specific queries can be viewed as a form of sense tagging, the aim of which is to reduce the amount of data to be browsed manually by the lexicographer.

6. TRANSLATION SELECTION AS WORD SENSE DISAMBIGUATION

It was pointed out at the beginning of this paper that, as Kilgarriff (1997) puts it, word senses can only be defined relative to a set of interests, which means that the granularity of word sense distinctions depends on the type of NLP application the developer has in mind. The first sections of this paper described several attempts to automate the WSD process with monolingual resources. The work carried out with the Collins-Robert dictionary and described in section 5 also refers to basically monolingual disambiguation, even if the resource is a bilingual dictionary. More recently, however, bilingual MRDs have attracted a lot of attention because some researchers have realised that such resources could contribute significantly to the design of intelligent dictionary look-up programs. Since the perspective here is a situation where a reader of an on-line text in a foreign language selects an unknown word and is provided with the most appropriate translation, the translation selection process can be seen as a variant of WSD. Michiels (1998) and Dufour (1997, 1998) describe the lexical component of DEFI, an English-French dictionary look-up system which works as a text-dictionary matcher which tries to find the lexical database entry (and hence the translation) whose linguistic and metalinguistic information (part of speech, collocational restriction, domain label) best matches the elements found in the source text (Dufour 1998:80).



The bilingual lexical database used by DEFI is a huge database resulting from the combination of two distinct bilingual dictionaries, the Collins-Robert E-F dictionary (Duval & Sinclair 1993, 3rd edition), and the Oxford-Hachette dictionary (Corréard & Grundy 1994). A detailed description of this very complex merging process, which combined information from the 2 sources into an enriched, machine-tractable database, can be found in Dufour (1997). The interest of merging the two dictionaries is obvious if one considers the following two entries taken respectively from the CR and OH dictionaries:

CR: **school** *n [fish]* banc
OH: **school** n **5** [of whales, dolphins, porpoises] banc

The resulting entry combines the lists of collocates for a given headword related to a given translation. The unification of the two lists yields an extended set of collocates:

school n [fish / whale / dolphin / porpoise] banc

When confronted with a phrase such as *a school of whales*, the DEFI user who clicks on *school* is provided with the appropriate translation (*banc*) because the system is able to recognise the tight collocational link which unites the two nouns. It should be noted that the DEFI matcher would have failed to select the correct translation if only the Collins-Robert dictionary had been used, since whales are mammals and CR only mentions *fish* as a typical collocate. Similarly, the phrase *a school of fish* would not be processed correctly if the Oxford-Hachette only had been used, since the only typical collocates listed in this dictionary are mammals and no match between the word *fish* in the source text and *whale/dolphin/porpoise* in the dictionary could ever be achieved.

Interestingly, the DEFI approach makes use of all relevant information in dictionary entries and resorts to other types of lexical resources in order to find the best match (and hence the most likely translation) even if the source text diverges from what can be found in the bilingual dictionaries. If only CR and OH were used, *a school of sardines* would never be interpreted appropriately and the 'default' French translation of *school*, *école*, would be proposed. This is due to the fact that the item *fish*, which appears as a collocate of *school*, should not be taken as the noun *fish* only, but also as the head of a thesauric class. For obvious space reasons, lexicographers are often not able to specify all possible combinations and frequently resort to

this practical device. To cope with this crucial problem, the DEFI developers have integrated two thesauri, Roget's thesaurus and WordNet, and use these additional resources to look for relevant hyperonymy or hyponymy relations. A user who clicks on *school* when reading the phrase *a school of sardines* triggers a mechanism which tries to establish a match between *school* and *sardine* in the DEFI bilingual database. In the absence of such a link, taxonomical hierarchies and thesaural classes are scanned in WordNet and in Roget's and, since *sardine* appears as a hyponym of *fish* and *fish* is linked to *school* in the dictionary, a match can be established. Dufour (1998) shows how the link between the words is granted a score, chosen empirically, depending on the kind of relations between them. Synonyms found in WordNet fare better than hypernyms, for instance, and the number of levels which separate the collocate found in the source text and the collocate found in the DEFI dictionary is also taken into account. For this reason, WordNet is sometimes too fine-grained and forces the matcher to move up (or down) more levels than should be necessary before finding an appropriate match, which means that lower scores are granted. The WordNet taxonomy for *sardine* looks as follows, for instance:

Sense 2

sardine -- (any of various small edible herring or related food fishes frequently canned)
=> clupeid fish, clupeid -- (any of numerous soft-finned schooling food fishes of shallow waters of northern seas)
=> soft-finned fish, malacopterygian -- (any fish of the superorder Malacopterygii)
=> teleost fish, teleost, teleostan -- (a bony fish of the subclass Teleostei)
=> bony fish -- (any fish of the class Osteichthyes)
=> fish -- (any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills)
=> aquatic vertebrate -- (animal living wholly or chiefly in or on water)
=> vertebrate, craniate -- (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
=> chordate -- (any animal of the phylum Chordata having a notochord or spinal column)
=> animal, animate being, beast, brute, creature, fauna -- (a living organism characterized by voluntary movement)
=> life form, organism, being, living thing -- (any living entity)
=> entity, something -- (anything having existence (living or nonliving))

For WSD, folk taxonomies (a sardine is_a fish, possibly even a whale/dolphin is_a type of fish) might prove more useful than fine-grained, scientific hierarchies such as found in WordNet. It would also be worth examining to what extent the performance of this approach could be enhanced if one took the familiarity indices into account and hence skipped the less familiar synsets and the less technical terms in moving up or down the hierarchies (see Tengi 1998:112-114).

In addition to external thesaural resources, the DEFI matcher also exploits internal information disseminated throughout the bilingual lexical database. In particular, the sharing of metalinguistic slots is used as a basis for measuring the semantic distance between the collocates in the source text and the collocates listed in the database.

It should also be noted that DEFI pays a lot of attention to the identification of multi-word units, since these are likely to pose problems to the prototypical user. It is now generally recognised that idiomatic phrases admit of a lot of syntactic manipulation and lexical variation (Moon 1998), which means that the recognition process must be very flexible. To tackle this problem, DEFI uses the same parser to analyse the source text and the dictionary examples and uses a number of heuristics in order to match the structures and the elements contained in the

two analyses. Michiels (1998) points out that the match is not an *either/or* decision, but a cline, especially since linguistic creativity is difficult to account for in standard dictionaries (see also Bauer *et al.* (1994) or, more recently, Aimelet *et al.* (1999) for a similar project using bilingual MRDs in dictionary look-up programs for translation selection and multi-word unit recognition).

7. CONCLUSION

The aim of this paper was not to produce an exhaustive survey of current approaches to semantic tagging. It deliberately left out a number of statistics-based approaches, rather focusing on the use of lexical resources, in particular machine-readable dictionaries and lexical databases. As is pointed out by Kilgarriff (1998), there are now a few WSD programs (at least for English), but large-scale resources usable in this perspective are still in their infancy. I have also deliberately left aside the question of determining which WSD program or semantic tagger is best, which is currently a hot topic in lexical semantics (see the recent SENSEVAL evaluation exercise coordinated by Kilgarriff in 1998). What researchers should bear in mind, of course, is that the tagging process is in fact based on a number of assumptions which may be justified from an engineering point of view: taggers are developed and senses are assigned as if words were divisible into discrete word senses, while it has been shown that a more appropriate approach would be to consider that distinct word senses result and emerge from the clustering of corpus citations. Such senses can only be arrived at after a close scrutiny of concordances excerpted from large, balanced corpora. In a nutshell, words shall be known by the company they keep (Firth 1959). The various projects and resources described in this paper are in a way all based on this motto and attempt to provide a partial solution to this key issue. Other efforts will undoubtedly be necessary.

8. REFERENCES

Aimelet, E., Brun, C., Griot, L.,& Segond, F. (1999): "A dictionary-based architecture for Word Sense Disambiguation", submitted to *ACL'99* (Association for Computational Linguistics).

Amsler, R.A. (1980): *The Structure of the Merriam-Webster Pocket Dictionary*, Ph.D. Thesis, University of Texas at Austin, Austin.

Atkins, B.T.S. (1992): "Tools for Computer-Aided Lexicography: The Hector Project", in *Acta Linguistica Hungarica*, Vol. 41 (1-4), pp. 5-71.

Atkins, B.T. & Duval, A. (1978): *Robert & Collins Dictionnaire Français-Anglais, Anglais-Français*, Paris: Le Robert/Glasgow: Collins. (3rd Edition edited by Sinclair, L. & Duval, A.).

Baker, C., Fillmore, C. & Lowe, J.B. (1998): "The Berkeley FrameNet Project", in *Proceedings of ACL/COLING 1998*.

Bauer, D., Segond, F. & Zaenen, A. (1994): "Enriching an SGML-Tagged Bilingual Dictionary for Machine-Aided Comprehension", Rank Xerox Research Centre Technical Report, MLTT, 11, Meylan.

Church, K., Gale, W., Hanks, P., Hindle, D. & Moon, R. (1994): "Lexical Substitutability", in Atkins & Zampolli (eds) *Computational Approaches to the Lexicon*, Oxford University Press, pp.153-177.

Church, K. & Hanks, P. (1990): "Word Association Norms, Mutual Information and Lexicography", in *Computational Linguistics*, 16 (3), pp.22-29.

Clear, J. (1994): "I Can't see the sense in a large corpus", in Kiefer, F., Kiss, G. & Pajzs, J. (eds) *Papers in Computational Lexicography COMPLEX'94*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, pp.33-47.

Corréard, M-H. & Grundy, V. (eds) (1994): *The Oxford-Hachette French Dictionary (French-English, English-French)*, Hachette and Oxford University Press.

Dufour, N. (1997): "DEFIDIC, a lexical database for computerized translation selection", in *Revue, Informatique et Statistiques dans les Sciences humaines*, N°1-4, CIPL, Liège: 79-111.

Dufour, N. (1998): "Recognizing collocational constraints for translation selection: DEFI's combined approach", in *EURALEX'98 Proceedings* - 8th International Congress of the European Association for Lexicography, University of Liège, Liège: 109-118.

Fellbaum, C. (ed.) (1998): *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, London, UK.

Fillmore, C. & Atkins, B.T.S. (1994): "Starting where the dictionaries stop: the challenge for computational lexicography", in Atkins, B.T.S. & Zampolli, A. (eds) *Computational Approaches to the Lexicon*, Oxford University Press, pp.349-393.

Fillmore, C. & Atkins, B.T.S. (1998): "FrameNet and lexicographic relevance", in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 28-29 May 1988.

Firth, J.R. (1959): *Selected Papers of J.R. Firth 1952-1959*, edited by F.R. Palmer, London and Harlow, Longman's Linguistics Library.

Fontenelle, Th. (1997a): *Turning a bilingual dictionary into a lexical-semantic database*, Lexicographica Series Maior 79, Max Niemeyer Verlag, Tübingen.

Fontenelle, Th. (1997b). "Using a bilingual dictionary to create semantic networks", in *International Journal of Lexicography*, 10 (4), pp.275-303.

Fontenelle, Th. (1998): "The semantic analysis of *of*-phrases for word sense disambiguation", in *EURALEX'98 Proceedings* - 8th International Congress of the European Association for Lexicography, University of Liège, Liège, pp.141-150.

Grefenstette, G. (1996). "Approximate Linguistics", in *Proceedings of the 4th Conference on Computational Lexicography and Text Research - COMPLEX'96*, Budapest, Hungary, Sept.1996.

Grefenstette, G. (1998): "The future of Linguistics and Lexicographers: Will there be lexicographers in the year 3000?", in *EURALEX'98 Proceedings* - 8th International Congress of the European Association for Lexicography, University of Liège, Liège, pp.25-41.

Grefenstette, G., Heid, U., Schulze, B.M., Fontenelle, T. & Gérardy, C. (1996): "Multilingual Collocation Extraction: The DECIDE project", in *EURALEX'96 Proceedings*, University of Göteborg, pp.93-107.

Harley, A. & Glennon, D. (1997): "Sense Tagging in Action", paper presented at the *ACL 1997 Conference on Tagging Text with Lexical Semantics: Why, What and How?*

Ide, N. & Véronis, J. (1998): "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", in *Computational Linguistics*, 24 (1), pp.1-40.

Kilgarriff, A. (1997): "I don't believe in word senses", in *Computers and the Humanities*, 31 (2), pp. 91-113.

Kilgarriff, A. (1998): "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs", in *EURALEX'98 Proceedings* - 8th International Congress of the European Association for Lexicography, University of Liège, Liège, pp.167-174.

Landes, S., Leacock, C., & Tengi, R. (1998): "Building Semantic Concordances", in Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA, London, UK, pp.199-216.

Leacock, C., Miller, G. & Chodorow, M. (1998): "Using Corpus Statistics and WordNet Relations for Sense Identification", in *Computational Linguistics*, 24 (1), pp.147-165.

Lowe, J.B., Baker, C. & Fillmore, C. (1997): "A Frame-Semantic Approach to Semantic Annotation", in *Tagging Text with Lexical Semantics: Why, What, and How? – Proceedings of the Workshop*, Special Interest Group on the Lexicon, Association for Computational Linguistics, pp.18-24.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.N., Mantha, S., Lessard, A. (1984/1988/1992): *Dictionnaire Explicatif et Combinatoire du Français Contemporain: Recherches Lexico-Sémantiques I, II, III*, Les Presses de l'Université de Montréal, Montréal.

Michiels, A. (1982): *Exploiting a Large Dictionary Database*. PhD Thesis, University of Liège, mimeographed.

Michiels, A. (1998): "The DEFI matcher", in *EURALEX'98 Proceedings* - 8th International Congress of the European Association for Lexicography, University of Liège, Liège, pp.203-211.

Michiels, A. & Dufour, N. (1996): "From SGML tapes to DIC clauses: Identifying Multi-Word Units for Context-Sensitive Lookup", DEFI Technical Report, Liège, available from the following URL: http://engdep1.philo.ulg.ac.be/michiels/defi.htm.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990): "Introduction to WordNet: An On-Line Lexical Database", in *International Journal of Lexicography*, 3 (4), pp.235- 244.

Moon, R. (1998): *Fixed Expressions and Idioms in English: A Corpus-Based Approach*, Oxford Studies in Lexicography and Lexicology, Oxford University Press, Oxford.

Procter, P. (ed.) (1978): *Longman Dictionary of Contemporary English*, (2nd edition edited by D. Summers), Longman Group Ltd, Harlow.

Procter, P. (ed.) (1995): *Cambridge International Dictionary of English*, Cambridge University Press.

Pustejovsky, J. (1995): *The Generative Lexicon*, The MIT Press, Cambridge, MA.

Schulze, B.M. & Christ, O. (1994): The CQP User's Manual. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Version 1.0d, May 1994 (Revised October 1994).

Schulze, B. (1996): *Macroprocessor User Guide*, IMS Universität Stuttgart, Internal Document.

Sinclair, J. (ed) (1987): *Collins COBUILD English Language Dictionary*, 1st Edition, HarperCollins, Glasgow.

Tengi, R.I. (1998): "Design and Implementation of the WordNet Lexical Database and Searching Software", in Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA, London, UK, pp.105-127.

Wilks, Y., Slator, B. & Guthrie, L. (1996): *Electric Words - Dictionaries, Computers and Meanings*, The MIT Press, Cambridge, MA and London.

Automatic Classification of Technical Terms using the NC-*value* Method for Term Recognition

KATERINA T. FRANTZI – SOPHIA ANANIADOU – JUNICHI TSSUJII

Abstract

Automatic term recogntion (ATR) has many applications in areas such as information retrieval and extraction from the web, summarisation, machine translation, dictionary construction etc. Aautomatic term classification is the grouping of terms into sets whose elements share conceptual properties. The combination of ATR and classification is important for structuring the acquired knowledge. Classes of terms can be used for special language (SP) dictionaries and thesauruses, the classification of documents, or the improval or precision or recall when searching for SP documents in the internet.

In this paper we present the *C-value/NC-value* a method for the automatic extraction of terms, and its further use for the recognition of similarities between them, to be used for their classification.

1. Introduction

Technical terms (henceforth called simply terms), are the linguistic realisation of specialised concepts, (Sager, Dungworth and McDonald 1980). Rapid changes in many specialised knowledge domains (particularly in areas like computer science, engineering, medicine etc.), means that new terms are being created all the time, making important the automation of their retrieval and classification.

Applications of ATR include information retrieval and extraction in digital libraries, the WWW, machine translation, summarisation, etc. In this paper we are particularly interested in multi-word terms. Many techniques for multi-word ATR move lately from using only linguistic information (Ananiadou 1988, Ananiadou 1994, Bourigault 1992), to incorporating

statistical as well. Dagan and Church, (Dagan and Church 1995), Daille et al., (Daille, Gaussier and Langé 1995), Justeson and Katz, (Justeson and Katz 1995), and Enguehard and Pantera, (Enguaehard and Pantera 1994), use frequency of occurrence. Daille et al., and Lauriston, (Lauriston 1996) also propose the likelihood ratio for terms consisting of two words. For the same type of terms, Damerau, (Damerau 1993), proposed a measure based on mutual information. Those of the above methods that aim to multi-word terms which may consist of more than two words, use as the only statistical parameter the frequency of occurrence of the candidate term in the corpus. A detailed description and evaluation of previous work on multi-word ATR can be found in (Kageura and Umino 1997, Frantzi 1998).

Automatic term classification attracts recently the interest of researchers, (Aizawa and Kageura 1998, Morimoto and Aizono and Kaji 1998, Vossen and Bloksma, 1998, Peters and Peters and Vossen, 1998), following the steps of ATR. The combination of ATR and classification is important for structuring the acquired knowledge. By automatic term classification we mean the grouping of terms into sets whose elements share conceptual properties. Working specifically on classifying terms rather than general language words can improve the processing time of the task, especially when terms and not general language words are needed, e.g. in areas like the classification of SP documents, or the improval or precision or recall when searching for SP documents in the internet.

Section 2 briefly presents *C-value* and section 3 the use of context information using*NC-value*. In this paper *NC-value* is not only used for the extraction of terms, but for the recognition of similarities between the extracted terms, as presented in section 4.

2. C-value

C-value is a domain-independent method for multi-word ATR which aims to improve the extraction of *nested terms*. The method takes as input a special language corpus and produces a list of candidate multi-word terms. These are ordered by their *termhood*. The output list is evaluated by a domain expert. Since the candidate terms are ranked according to their termhood, the domain expert can scan the lists starting from the top, and go as far down the list as time/money allow.

The *C-value* approach combines linguistic and statistical information, emphasis being placed on the statistical part. The linguistic information consists of the part-of-speech tagging of the corpus, the linguistic filter constraining the type of terms extracted, and the stop-list. The statistical part combines statistical features of the candidate string, in a form of measure.

We use Brill's part-of-speech tagger, (Brill, 1992), and a stop-list that consists of 229 function and other content words, picked from a sample of our corpus (1/10). The words that are included in the stop-list exhibited high frequencies in that sample of the corpus. Some examples are: *great, numerous, several, year, just, good*, etc. Regarding the linguistic filter, since most terms consist of nouns and adjectives, (Sager, 1990), and sometimes prepositions, (Justeson and Katz 1995), we use linguistic filters that accept these types of terms. Since different applications require different filters, we test the method with each of the 3 filters, (Frantzi 1998, Frantzi and Ananiadou 1998):
1.$NounNoun^{+}$,
2.$(Adj|Noun)^{+}Noun$,
3.$((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)^{?})(Adj|Noun)^{*})Noun$,
In this paper we use the results produced using the second of the above filters.

The *C-value* statistical measure assigns a termhood to a candidate string, ranking it in the output list of candidate terms. The measure is built using statistical characteristics of

the candidate string. These are:
1.The total frequency of occurrence of the candidate string in the corpus.
2.The frequency of the candidate string as part of other longer candidate terms.
3.The number of these longer candidate terms.
4.The length of the candidate string (in number of words).

Since the maximum length terms can not be nested in longer terms, and some strings are never found as nested anyway, we distinguish two cases

1. If a is a string of maximum length or has not been found as nested, then its termhood will be the result of its total frequency in the corpus and its length.

2. If a is a string of any other shorter length, then we must consider if it is part of any longer candidate terms. If it appears as part of longer candidate terms, then its termhood will also consider its frequency as a nested string, as well as the number of these longer candidate terms. Though the fact that it appears as part of longer candidate terms affects its termhood negatively, the bigger the number of these candidate terms, the higher would be its independence from these. This latter number moderates the negative effect of the candidate string being nested in longer candidate terms.

The measure of termhood, called *C-value* is given as

$$C\text{-}value(a) = \begin{cases} log_2|a| \cdot f(a) & a \text{ is not nested,} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & otherwise \end{cases} \quad (1)$$

where
a is the candidate string,
$f(.)$ is its frequency of occurrence in the corpus,
T_a is the set of extracted candidate terms that contain a,
$P(T_a)$ is the number of these candidate terms.
More on *C-value* can be found in (Frantzi 1998, Frantzi and Ananiadou 1998).

3. Context Information

The environment of a word is often used to identify or clarify its meaning. In automatic systems the information used for disambiguation is restricted mainly to surface criteria as opposed to semantic, discourse and pragmatic information. Lexical information from the context of words has been used for the construction of thesaurus dictionaries, (Grefenstette 1994). In that case, the context of a word provides clues to its meaning and its synonyms. Grefenstette's system, SEXTANT, uses local lexical information to acquire synonyms. Words that are used in a lexically similar way are candidates to be synonymous. The nouns, adjectives and verbs from the context of the examined word are used to give hints for its meaning.

Regarding term recognition, Sager et al., (Sager and Dungworth and McDonald 1978), stated that multi-word terms differ from extended word units by that they cannot be freely modified, and can only accept a limited number of qualifiers.

Since extended term units differ from extended word units as far as modification is concerned, we could use information from the modifiers to distinguish between terms and non-terms. Thus, if *consistent* is an adjective that tends to precede terms in medical corpora, and it occurs before a candidate term string, we could exploit this information for the benefit of term recognition. Besides adjectives and nouns, we can expand the use of modifier types to verbs that belong to the environment of the candidate term: the string *show* of the verb *to*

show in medical domains is often followed by a term, e.g. *shows a basal cell carcinoma.* The string *called* of the verb *to call*, and the form *known* of the verb *to know*, are often involved in definitions, e.g. *is known as the singular existential quantifier* and *is called the Cartesian product.* We will use the three part-of-speech elements also used by (Grefenstette 1994 to obtain information about the termhood of a candidate string, when they either precede or follow it. These are:
1.nouns (*compound cellular naevus*),
2.adjectives (*blood vessels are present*), and
3.verbs (*composed of basaloid papillae*).

Now, let as describe a way to create a list of 'important' *term context words* from a set of terms extracted from a specialised corpus. By term context words we mean those that appear with terms in texts. The context words we treat are adjectives, nouns and verbs that either precede or follow the candidate term.

The criterion for the extraction of a word as a term context word is *the number of terms it appears with.* The assumption is that the higher this number, the higher the likelihood that the word is 'related' to terms, and that it will occur with other terms in the same corpus. Term context words for a specific domain/corpus are not necessarily the same for another domain/corpus. The words *present, shows, appear, composed* tend to appear with terms in our medical corpus. The measure to give weight to term context words is given by

$$Weight(w) = \frac{t(w)}{n} \tag{2}$$

where
w is the context word (noun, verb or adjective) to be assigned a weight as a term context word,
$Weight(w)$ the assigned weight to the word w,
$t(w)$ the number of terms the word w appears with,
n the total number of terms considered.
The term context words can be now ranked according their weight.

NC-value is the method which incorporates context information into the *C-value* method for the extraction of multi-word terms. Assuming we have a corpus from which we want to extract the terms, we have three stages.
First stage:
We apply the *C-value* method to the corpus. The output of this process is a list of candidate terms, ordered by their *C-value.*
Second stage:
This involves the extraction of the term context words and their weights. These will be used in the third stage to improve the term distribution in the extracted list. In order to extract the term context words, we need a set of terms, as describted before. We have chosen to keep the method domain-independent and fully-automatic at this stage, therefore we do not use any external source (e.g. a dictionary) to provide us with the set of terms to be used for this purpose. We use instead the 'top' candidate terms from the *C-value* list since these present high precision on real terms, and tolerate the 'noise' produced by the few non-terms found on the top of the list. These 'top' terms produce a list of term context words, which are assigned a weight as described.
Third stage:
This involves the incorporation of context information acquired from the second stage of the extraction of multi-word terms. The *C-value* list of candidate terms extracted during stage

one is re-ranked using context information, so that the real terms appear closer to the top of the list than they did before, i.e. the concentration of real terms at the top of the list increases while the concentration of those at the bottom decreases. The re-ranking takes place in the following way:
Each candidate term from the *C-value* list appears in the corpus with a set of context words. From these context words, we retain the nouns, adjectives and verbs for each candidate term. These words may or may not have been met before, during the second stage of the creation of the list with the term context words. In the case where they have been met, they retain their assigned weight. Otherwise, they are assigned zero weight. For each candidate term, we obtain the context factor by summing up the weights for its term context words, multiplied by their frequency appearing with this candidate term. This is the second factor of the *NC-value* measure, the first being *C-value* itself. The *NC-value* measure is given by

$$NC\text{-}value(a) = 0.8\,C\text{-}value\ (a) + 0.2 \sum_{b \epsilon C_a} f_a(b) weight(b) \tag{3}$$

where
a is the candidate term,
C_a is the set of distinct context words of a,
b is a word from C_a,
$f_a(b)$ is the frequency of b as a term context word of a,
$weight(b)$ is the weight of b as a term context word.

The two factors of *NC-value*, i.e. *C-value* and the context information factor, have been assigned the weights 0.8 and 0.2 respectively. These have been chosen among others after experiments and comparisons of the results, as we will discuss in the following section.

Table 3 gives the first 40 candidate terms extracted by *NC-value*. Column 1 gives the *NC-value* of the candidate term, column 2 its *C-value* and column 3 its frequency of occurrence. More on *NC-value* can be found in (Frantzi 1998, Frantzi and Ananiadou 1998).

4. Similarity Calculations for Term Classification

NC-value provides us with the context words for each term extracted. As context words we will use the nouns, adjectives and verbs of the term's environment.

As an example, table 1 gives context words that appear with the extracted terms *central retinal vein, ciliary choroidal melanoma, cystic basal cell* and *fibrous scar tissue*. Column 1 gives the frequency with which the context word of column 2 appears with the term. Column 3 gives the weight of this context word.

Having this context information, we apply *Jaccard Similarity Measure*, (Grefenstette 1994). The binary Jaccard measure for two objects m and n is given by

$$\frac{\text{Count(Attributes shared by } m \text{ and } n)}{\text{Count(Unique attributes possessed by } m \text{ or } n)} \tag{4}$$

In our case the objects m and n are the terms to be compared for similarity. The attributes are their context words in the corpus.

In order to get the extracted terms classified we will apply a comparison between terms, like the one in (Grefenstette 1994) between nouns, using Jaccard measure:

for each term m extracted
    for each other term n

frequency	context word	weight
Context words appearing with *central retinal vein*		
1	APPEARS	0.0357143
1	AREA	0.0357143
19	ARTERY	0.0214286
1	COMPOSED	0.214286
1	COMPRISING	0.0714286
2	FOLLOWING	0.107143
1	FORM	0.171429
1	OLD	0.107143
1	OLD	0.107143
Context words appearing with *ciliary choroidal melanoma*		
1	ARTERY	0.0214286
12	BODY	0.121429
1	DETACHMENT	0.0642857
1	DIAGNOSIS	0.0142857
1	EXTENSIVE	0.142857
1	IS	0.435714
2	SPACE	0.0428571
Context words appearing with *cystic basal cell*		
1	ABNOID	0.0142857
31	ADENOID	0.05
1	ARISING	0.214286
1	CONTAIN	0.185714
1	CONTAINING	0.392857
1	EXTENSIVE	0.142857
1	FEW	0.207143
3	FORM	0.171429
1	FORMING	0.157143
3	IS	0.435714
6	LARGE	0.0285714
1	LEFT	0.242857
1	SEEN	0.121429
2	SHOWING	0.257143
1	SMALL	0.214286
2	SOLID	0.1

Table 1: Context words of three extracted terms.

use Jaccard's measure to calculate the similarity between m and n
sort the terms according to their similarity with m
keep the "most" similar to m

As an example, let us consider the terms
hyaline fibrous pannus
hyaline fibrous plaque
optic nerve head

optic nerve vitreous
intra occular pressure
fibro vascular tissue
We will calculate the similarity between every two of them using the binary Jaccard measure. Table 2 gives the number of different context words that each of the above terms appears with. If by $Sim(x, y)$ we mean the similarity between terms x and y, then according to the

Term	no. of context words
hyaline fibrous pannus	51
hyaline fibrous plaque	49
optic nerve head	111
optic nerve vitreous	97
intra occular pressure	34
fibro vascular tissue	71

Table 2: Terms and the number of their context words

binary Jaccard measure,

$$Sim(\text{hyaline fibrous pannus,hyaline fibrous plaque}) = \frac{49}{51 + 49 - 49} = 0.96 \qquad (5)$$

Sim(hyaline fibrous pannus,optic nerve head) = 0.1095
Sim(hyaline fibrous pannus,optic nerve vitreous) = 0.1212
Sim(hyaline fibrous pannus,intra occular pressure) = 0.0759
Sim(hyaline fibrous pannus,fibro vascular tissue) = 0.2708

Sim(hyaline fibrous plaque,optic nerve head) = 0.1112
Sim(hyaline fibrous plaque,optic nerve vitreous) = 0.123
Sim(hyaline fibrous plaque,intra occular pressure) = 0.0779
Sim(hyaline fibrous plaque,fibro vascular tissue) = 0.2765

Sim(optic nerve head,optic nerve vitreous) = 0.8738
Sim(optic nerve head,intra occular pressure) = 0.0719
Sim(optic nerve head,fibro vascular tissue) = 0.103

Sim(optic nerve vitreous,intra occular pressure) = 0.0826
Sim(optic nerve vitreous,fibro vascular tissue) = 0.1125

Sim(intra occular pressure,fibro vascular tissue) = 0.05

We can observe that regarding their common context, *hyaline fibrous pannus* is much more similar to *hyaline fibrous plaque* than to *intra occular pressure*. That *optic nerve head* is much more similar to *optic nerve vitreous* than to *fibro vascular tissue*, etc.

With this kind of calculations, for each term we can create a list with its "similar" terms in decreasing order of importance, and then terms that show a great degree of similarity can be put in the same set. We have yet to implement this.

Our next experiments will also involve the use of the *weighted* Jaccard measure instead the binary one. This is given by

$$\frac{\sum_{unique_attributes} min(weight(object_m, attribute), weight(object_n, attribute)}{\sum_{unique_attributes} max(weight(object_m, attribute), weight(object_n, attribute)} \quad (6)$$

The weights in the formula above would be the weights that have been assigned to the context words according to their importance as term context words, as described in section 3.

5. Summary and Future Work

This paper presented the *C-value/NC-value* domain-independent method for the semi-automatic extraction of multi-word terms from special language English corpora. It showed how *NC-value* can be used for the identification of similarities between the extracted terms. These similarities will be used for the automatic classification of the terms.

Regarding direct future work, the implementation for the creation of the term sets has yet to be completed and the semantic/conceptual relations have to be established.
The evaluation will involve the use of existing hierarchies and thesauruses and the judgment of domain experts. Following this, future directions will be on the applications for the produced term sets.

References

[1] Aizawa, A.N., Kageura, K.:. An Approach to the Automatic Generation of Multilingual Keyword Clusters. Proceedings of the 1st Workshop on Computational Terminology, Computerm'98, 17th International Conference on Computational Linguistics, COLING-ACL'98, (1998) 8-14

[2] Ananiadou, S.: Towards a Methodology for Automatic Term Recognition. Ph.D. Thesis, University of Manchester Institute of Science and Technology (1988)

[3] Ananiadou, S.: A Methodology for Automatic Term Recognition. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, (1994) 1034–1038

[4] Bourigault, D.: Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. Proceedings of the 14th International Conference on Computational Lingustics, COLING'92, (1992) 977–981

[5] Brill, E.: A simple rule-based part of speech tagger. Proceedings of the 3rd Conference of Applied Natural Language Processing, ANLP'92, (1992)

[6] Dagan, I., Church, K.: Termight: Identifying and Translating Technical Terminology. Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL'95, (1995) 34–40

[7] Daille, B., Gaussier, E., Langé, J.: Towards Automatic extraction of Monolingual and Bilingual Terminology. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, (1994) 515–521

[8] Damerau, F.J.: Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. Information Processing & Management **29** (1993) 433–447

[9] Enguehard, C., Pantera, L.: Automatic Natural Acquisition of a Terminology. Journal of Quantitative Linguistics **2** (1994) 27–32

[10] Frantzi, K.T.: Automatic Recognition of Multi-Word Terms. Ph.D. Thesis, Manchester Metropolitan University Dept. Of Computing & Mathematics, in collaboration with UMIST Centre for Computational Linguistics, (1998)

[11] Frantzi, K.T., Ananiadou, S., Tsujii, J.: The *C-value/NC-value* method of Automatic Recognition for Multi-Word Terms. In Lecture Notes in Computer Science, 1513, Springer-Verlag, (1998) 585–604

[12] Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, (1994)

[13] Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering **1** (1995) 9–27

[14] Kageura, K., Umino, B,: Methods of Automatic Term Recognition -A Review-. Terminology **3** (1996) 259–289

[15] Lauriston, A.: Automatic Term Recognition: performance of Linguistic and Statistical Techniques. Ph.D. Thesis, University of Manchester Institute of Science and Technology (1996)

[16] Morimoto, Y., Aizono, T., Kaji, H.: Generation of a Corpus-dependent Thesaurus and Interactive Text Retrieval. Proceedings of the JSPS–Hitachi Workshop on New Challenges in NLP and its Applications, edited by J. Tsujii and H. Kaji, (1998), 65–68

[17] Peter, W., Peters, I., Vossen, P.: Building Consistent Terminologies. Proceedings of the 1st LREC, (1998) 409–416

[18] Sager, J.C.: Commentary by Prof. Juan Carlos Sager, Actes Table Ronde sur les Problèmes du Découpage du Terms, Montréal, 26 aouũt. Guy Rondeau, AILA–Comterm,Office de la Langue Francaise, Québec, (1978) 39–74

[19] Sager, J.C., Dungworth, D., McDonald, P.F.: English Special Languages: principles and practice in science and technology. Oscar Brandstetter Verlag KG, Wiesbaden, (1980)

[20] Sager, J.C.: A Practical Course in Terminology Processing. John Benjamins Publishing Company, (1990)

[21] Vossen, P., Bloksma, L.: Categories and Classifications in EuroWordNet. Proceedings of the 1st LREC, (1998) 399–407

NC-value	*C-value*	*frequency*	candidate term	real term?
1652.75	2025.41	2084	OPTIC NERVE	1
1328.8	1656.53	1666	DESCEMET'S MEMBRANE	1
1257.39	1544.63	984	BASAL CELL CARCINOMA	1
1181.83	1449.47	1538	BASAL CELL	1
1121.73	1362.63	1387	FIBROUS TISSUE	1
1107.11	1377	1377	PLANE OF SECTION	0
1015.74	1210.57	1214	ANTERIOR CHAMBER	1
874.577	1092	1102	CORNEAL DIAMETERS	1
863.693	1077	1084	BOWMAN'S MEMBRANE	1
760.691	936.917	1025	CELL CARCINOMA	1
745.908	931.958	592	STUMP OF OPTIC NERVE	0
707.082	876.667	882	PLASMA CELLS	1
608.836	759.197	484	BASAL CELL PAPILLOMA	1
589.389	733.333	741	MALIGNANT MELANOMA	1
534.012	3	3	T CELL	1
528.179	658	658	NASAL SIDE	1
513.398	622.89	400	HYALINE FIBROUS TISSUE	1
498.97	623	623	TRABECULAR MESHWORK	1
486.845	598.846	621	LID MARGIN	1
464.21	576	597	CORNEAL DISC	1
386.869	476	534	NERVE HEAD	1
350.517	437	437	PLANE OF SECTION=	0
341.632	424.77	274	OPTIC NERVE HEAD	1
337.273	420.5	433	MELANOMA OF CHOROID	0
332.391	413	413	PLANES OF SECTION	0
323.011	398	407	AXIAL REGION	1
314.719	378	383	KERATINOUS CYST	1
304.634	379.5	506	CELL PAPILLOMA	1
296.572	366.429	370	CILIARY PROCESSES	1
293.046	364.5	373	BRUCH'S MEMBRANE	1
261.95	4	4	B CELL	1
256.436	318	329	ELLIPSE OF SKIN	0
250.812	307.483	197	CELLULAR FIBROUS TISSUE	1
247.383	298	305	LYMPHOCYTIC INFILTRATION	1
244.758	299.333	303	OCULAR STRUCTURES	1
234.196	285.091	295	LENS CAPSULE	1
227.84	284	284	SEBACEOUS CYST	1
216.064	269	278	PUPILLARY BORDER	1
210.917	244.5	249	CORNEAL EPITHELIUM	1
210.832	6	11	B CELLS	1
205.614	256.764	165	WEDGE OF LID MARGIN	0

Table 3: The first 40 candidate terms extracted by *NC-value*

SGML/XML tools

PÉTER HUSZÁR

Abstract: The aim of this paper is to describe available SGML/XML tools on the market. We evaluate the functionality, usability and price aspects of different categories of tools. These categories include DTD development utilities, parsers/validators, viewers, editors and data manipulators.

Foreword

The following is an overview of the widely used and/or supported products. It is not our aim to discuss each and every product in very deep details.
All product names mentioned herein are the trademarks of their respective owners.

DTD development tools

DTD development is mainly *analysis* and *design* which requires human skills and knowledge but no actual tools. There are books available on the market which helps you in this analysis [Maler/Andaloussi 1996], [Aschuler 1995].
The only serious tool on the market which helps you in DTD *coding* is Near&Far by Microstar Ltd. This is a graphical tool to visualize the tree-hierarchy of a DTD and enables visual, user-friendly creation and editing. Newest version 3.0 has the following features:

- both SGML and XML DTDs can be created
- SGML DTDs can be converted to XML DTDs and vica versa

This product is available on Windows platforms. Further information is available at
http://www.microstar.com/products.html

Parsers/Validators

Most of the products (editors, data manipulators, complex systems) contain a built-in parser/validator for checking SGML documents. But there are standalone tools to perform such a check independently from any other task.

SP/XP by James Clark: A validating parser. Strictly speaking SP is an event based programming library which reads the SGML document and gives back structured information and validating errors. It has an API that allows SP to be incorporated in other tools. Many products on the market use SP as a built-in parser. There are several user interfaces developed around the library. James Clark has NSGMLS (extended version of SGMLS), as runtime environment for SP. SP is freeware. XP is the XML version of SP. The package contains SX an SGML to XML converter and other utilities. Further information is available at
http://www.jclark.com/sp.html

SGML editors

The products discussed in this category has built-in SGML/XML interface and intelligence. They are capable of reading and understanding DTDs and helping the creation and editing of SGML/XML documents. Except where it is noted these products has no or limited formatting capabilities. They can only produce draft printout.

InContext V2.11 by Siemens Nixdorf Informationssysteme AG: A Windows based SGML editor. Unlike the other editors it emphasizes the tree-hierarchy and allows editing in forms. It uses the DTD in its original ASCII format (no DTD compilation is necessary). It has very limited entity-management capabilities and no programming interface. InContext is a part of the Corel Ventura package. It runs on Windows platforms.
Because of its very limited capabilities we suggest to use InContext only for very small and simple tasks. Further information is available at
http://www.sni.de/public/aswba/sdp/produkte/incontex/english/incontex.htm

Author/Editor 3.5 by Interleaf Inc.: Part of the Panorama SGML package. It has a friendly, visual user interface, a buit-in automatic validator which lets the user to create SGML valid documents and built-in templates for the most common document types (article, book, minute etc.). It uses the DTD in compiled binary format. For creating this binary format the user should use RulesBuilder 3.0, another product of the Panorama package. It has limited entity-management capabilities and no programming interface (Sculptor 1.0, another product of the Panorama package adds programming capabilities to Author/Editor). Available on several Windows and UNIX platforms and on Mac.
Author/Editor is a good choice for editing English language documents. Further information is available at

http://www.interleaf.com/Panorama/page2.html

WordPerfect 8 by Corel Inc.: The new versions of the popular word processor come with an SGML extension. Main features are the user friendly environment, the built-in validator and the macro language. It uses a compiled DTD but the compiler is the part of the package, too. It has very good styling capabilities. Uses its own stylesheet format. Some SGML features are not supported causing minor implementation problems in complex projects. We should mention its low price compared to other editors.
However WordPerfect is not a native SGML editor it gives you an above average environment for editing SGML documents. Further information is available at

http://www.corel.com/products/wordperfect/cwps8/index.htm

Adept Editor 8.0 by ArborText Inc.: One of the most sophisticated editors. After more than ten years of development Adept gives you the most functionality and the most support for SGML features. The newest version also supports XML. It has friendly interface and a programming language (ACL). It uses compiled DTDs which can be created by Document Architect (also by ArborText). Adept Editor supports FOSI stylesheets which are created and maintained also by Document Architect. Adept has full entity-management capabilities. Adept is available on Windows and Unix platforms and on OS2.
Adept is the choice of serious users for serious tasks. Further information is available at

http://www.arbortext.com/editor.html

PSGML: This is a major mode for emacs. PSGML has good SGML support and its free. It is not as friendly as the other editors but is a good choice for everyone who has experience with emacs and Unix-like working environments. Further information is available at

http://www.lysator.liu.se/projects/about_psgml.html

XML Editors

There are many XML editors under construction. They have 0.x or 1.x versions available. Most of them are free or available for low cost but their functionality is limited. The reader is encouraged to look after them on the WEB. Here we mention only one product of a company which has much experience on this field.

XMetaL 1.0 by SoftQuad Inc.: This product is based on HoTMetaL a popular and professional HTML editor. Now it supports the XML Recommendation with CSS styles. Contains templates, a scripting engine for JavaScript and VBScript, supports CALS table model, and has three editing views (wysiwyg, tagged and source). It is not free but a professional tool for XML projects. Further information is available at

http://www.softquad.com/products/xmetal

SGML Viewers

One of the main tasks in connection of an SGML document is to show it for human reading. Mostly editors serve for this purpose as they use a kind of stylesheet to show readable SGML document. But there are special SGML viewers also to perform this task. Again we mention only the most widely used product.

Panorama Viewer 2.0 by Interleaf Inc.: A part of the Panorama package. Allows Handling of any type of DTD. Uses SP as built-in parser. Panorama Viewer uses its own stylesheet that can be interactively modified. Supports many graphic and multimedia format. Basic version is free but professional version is not. Can be configured as a plug-in to HTML browsers. Further information is available at

http://www.interleaf.com/Panorama/page3.html

XML browsers

As XML is designed for the Internet, browsers will serve as XML viewers. Microsoft's Internet Explorer 5.0 already supports XML documents at the moment in a tree-like view. Netscape's Communicator 5.0 will also support XML when it will be published.
In the XML technology the stylesheet that governs visualization is standardized (XSL) but this standardization process is just not finished. It is expected that after XSL will be a recommendation both browsers will support it.

Data manipulators

There are products on the market to process SGML data. Their task is to convert non-SGML data to SGML or non-SGML data, transform SGML data to non-SGML or SGML data (in the latter case transform to different DTD) or manipulate the information of an SGML document (create summaries, statistical information, metadata etc.).

DynaTag by Inso: An interactive tool to convert Rich Text Format, Interleaf and FrameMaker files to a DTD using the Rainbow DTD as an intermediate result. The Rainbow DTD is a public domain DTD that captures the formatting aspects of these document types. Based on these aspects the user can create his/her own elements. DynaTag itself is a special purpose tool and is unable to solve all conversion problems. DynaTag mainly used for preconversion. Further information is available at

http://www.inso.com/dynatext/dtagbrief.htm

Balise 4.0.3 by AIS: A complete translation tool. Balise supports SGML, RTF, HTML and XML data. It has a procedural and an event based language both interpreted. There is a built-in parser, which can be replaced by custom parsers. Balise can handle SGML structure and context. It has a C and a C++ API. It has limited capabilities for data conversion. Balise has a runtime and a developer version. Further information is available at

http://www.balise.com

Omnimark 4 by Omnimark Technologies Corp.: A complete tool for conversion and translation including a built-in parser. It has its own scripting language that is interpreted by the program. Omnimark tracks the SGML hierarchy through the document and allows context-driven processing. Current version goes beyond this functionality. A light version is available for non-profit purposes. Further information is available at

http://www.omnimark.com/magazine/index.html

Summary

If you start with SGML you can choose from several tools depending on your needs and your experience. You can get free tools with limited functionality and support and you can get complex tools with full support for a bag of money.
XML is relatively new but already with many tools. XML is cheaper than SGML but the tools are not so sophisticated yet. Whenever you consider to use a tool you should check its SGML/XML capability and the functionality and support it gives. You can be sure: there are tools for every need.

References

[Maler/Andaloussi 1996] Eve Maler, Jeanne El Andaloussi, *Developing SGML DTDs*, Prentice Hall PTR, 1996.
[Aschuler 1995] Liora Aschuler, *ABCD...SGML*, International Thomsom Computer Press, 1995

English–Japanese Parallel Corpus Developed by JEIDA

HITOSHI ISAHARA

abstract

This paper describes the bilingual corpus project by JEIDA (Japan Electronics Industry Development Association). JEIDA decided to develop its own bilingual (English-Japanese) corpus for NLP research and make it publicly available without charge. The main purpose of this project is to develop a medium sized aligned parallel corpus of English and Japanese. Also through this project, we have the opportunity to discuss various facets involved in the development of a bilingual corpus, to do research on the alignment of Japanese and English sentences and to investigate automatic acquisition of linguistic knowledge using the developed corpus. The first version of our corpus has only SGML tagged text, sentence-level alignment data and manually added comments. We are trying to add more tags to our bilingual corpus in order to get more precise information, such as part-of-speech tags, word alignment tags, clause alignment tags and syntactic and semantic tags. Currently, we are adding phrase level alignment tags and correspondences between proper nouns and compound nouns in Japanese and English.

1. Introduction

A huge amount of bilingual data is necessary for NLP (Natural Language Processing) research, for example, corpus-based research on MT systems. It can be used to extract fundamental data for research and to verify the research results.

There are several levels of bilingual corpora. The standards are pairs of documents on the same subject written in two different languages. There are corpora with various kind of correspondences, at, for example, chapter, paragraph and sentence level. There might be corpora with correspondences of sentence structures or words in the sentences. The more precise the tagged corpus, the more useful it is for ordinary NLP research, such as the development of machine translation systems. However, in some cases, sets of bilingual documents which belong to the same topic domain but no have link between them are also useful for research in NLP. For example, because we can predict that two documents on a similar topic have a similar distribution of words, this kind of corpus is useful for information retrieval research. There are some such corpora for Indo-European languages. However, there are few such bilingual corpora for Japanese and other languages that are generally available for research purposes. This is due to the problem of copyright and the cost of data development.

When we think about a sentence-to-sentence aligned bilingual corpus, we have to distinguish between sentences with context and sentences without context. The corpus with sentence-level alignment with context is much more useful for a wider range of NLP research. So far, research relating Japanese and English has mostly been done using example sentences in the Kodansha's Japanese to English dictionary. However, example sentences in a dictionary are without context.

Even if researchers try to develop their own bilingual corpus, there are few huge bilingual resources, i.e., documents in both Japanese and English. We are using Japanese newspaper articles by Mainichi Shimbun for research on Japanese processing, however, there is no complete English translation of Japanese newspaper articles. Some newspaper companies in Japan issue newspapers in English. However, even if the events discussed are the same, the content of the Japanese version and English version differ, due to the difference in the targeted audience. To gather these pairs of newspaper articles might be useful for information retrieval research, as we described above, but not for ordinary NLP research. There are many Japanese translations of novels written in English. However, in order to collect them in bilingual corpora and make them publicly available, one must get permissions from authors, translators, and publishers in the foreign countries and from publishers in Japan. This is extremely difficult.

Therefore, JEIDA (Japan Electronics Industry Development Association) decided to develop its own bilingual (English-Japanese) corpus for NLP research and make it publicly available without charge. The main purpose of this project is to develop a medium sized aligned parallel corpus of English and Japanese. Also through this project, we have the opportunity to discuss various facets involved in the development of a bilingual corpus, to do research on the alignment of Japanese and English sentences and to investigate automatic acquisition of linguistic knowledge using the developed corpus.

JEIDA is a joint organization of computer-related companies in Japan. The committee on text processing technology is a subcommittee of JEIDA's natural language processing committee. This subcommittee has been developing their bilingual aligned corpus for research in NLP, since the 1996 Japanese fiscal year. In fiscal year 1996, we did a feasibility study and received permission from the Japanese Ministries to create such a resource. We, then, made a "small" sentence aligned corpus in fiscal year 1997. A new project started in April, 1998, was aimed at developing a much larger corpus with more precise tags. An overview of this bilingual corpus project is presented in this paper.

2. The Source of our Corpus

We first decided on the source documents. White papers from Japanese Ministries were selected for the following reasons:

(1) white papers are well edited and the quality of the language is high,
(2) both Japanese versions and their precise English translations exist,
(3) governmental papers have fewer copyright problems than commercial publications, and
(4) white papers cover a wide range of topics.

As for (1), sentences in the white papers are supposed to be written in a very specific style. Moreover, because they are edited several times before completion, the sentences are of a higher grammatical quality than, for example, sentences in newspaper articles.

Because of (2), the English sentences in these white papers would not be considered "good" contextual translations but are merely sentence-to-sentence or paragraph-to-paragraph translations. However, in this respect, they suit the current state of NLP research. Current NLP technologies are mainly used for processing sentences without context, not text as a whole.

As for (3), even if there is no copyright problem for the original Japanese texts, there might be some claims of rights for their translation by the translators.

We have already gotten permission to use white papers from three Japanese ministries: the Environment Agency, the Economic Planning Agency and the Science and Technology Agency. We have developed an aligned bilingual corpus using six white papers from the 1992 to 1996 fiscal years (Table 1) and are now enlarging it with six other white papers.

The size of each document is also shown in Table 1. There are more English sentences than Japanese sentences. That is because these pairs of texts are original Japanese sentences and their English translations. Translators sometimes translate one sentence in a source language into multiple sentences in a target language. They seldom translate several sentences into one sentence. Also, Japanese style of writing favors long, complex sentences while English style prefers shorter sentences.

Table 1: Source of aligned bilingual corpus

White Paper		Size (byte)	Section	Paragraph	Sentence
Environment (Heisei 6th)	Japanese	1,175k	693	2,100	4,525
Environment (1993-1994)	English	1,535k	693	2,238	6,432
Economic Planning (Heisei 7th)	Japanese	601k	332	1,291	3,080
Economic Planning (1994-1995)	English	741k	332	1,279	3,645
Economic Planning (Heisei 8th)	Japanese	520k	339	816	2,761
Economic Planning (1995-1996)	English	766k	339	824	3,265
Science and Technology (Heisei 6th)	Japanese	417k	289	948	1,738
Science and Technology (1994)	English	655k	289	1,307	2,471
Science and Technology (Heisei 7th)	Japanese	434k	326	967	1,881
Science and Technology (1995)	English	689k	326	1,277	2,695
Science and Technology (Heisei 8th)	Japanese	383k	254	828	1,630
Science and Technology (1996)	English	604k	254	944	2,375

We input texts into the computer and tagged them based on the TEI format. Details of these processes are described in the following section. Examples of the texts from the white paper of the Environment Agency are shown in Figure 1.

We are trying to get permission to use another kind of document, such as monthly journals and manuals. This would make our corpora more "balanced". We have just gotten permission from a software company to use their manuals in Japanese and in English for our corpus.

Japanese	English
<div4 type=subsection id="J2.1.1.4"> <head id="J2.1.1.4-h" corresp="E2.1.1.4-h"> (4) 農林水産物の生産と消費の増大 </head> <p id="J2.1.1.4-1" corresp="E2.1.1.4-1"> <s id="J2.1.1.4-1.1" corresp="E2.1.1.4-1.1 E2.1.1.4-1.2"> 農林水産業は、食糧や木材等の供給により人類の生存を最も基礎的なところで支えている重要な活動であり、また、農林水産業が営まれている地域においては、適切な農林水産活動を通じて農地、森林等が有する環境保全能力が維持されている。</s> <s id="J2.1.1.4-1.2" corresp="E2.1.1.4-1.3 E2.1.1.4-1.4"> 一方、その生産活動に伴い、途上国を中心に進んでいる森林から農地への改変といった資源利用や欧米諸国における肥料等の使用による水質汚濁、家畜等からの温室効果ガスの一種であるメタンガスの発生という形で、環境に負荷を与えている。</s> <div5 type=subsubsection id="J2.1.1.4.1"> <head id="J2.1.1.4.1-h" corresp="E2.1.1.4.1-h"> ア. 主食生産</head> <p id="J2.1.1.4.1-1" corresp="E2.1.1.4.1-1"> <s id="J2.1.1.4.1-1.1" corresp="E2.1.1.4.1-1.1 E2.1.1.4.1-1.2"> 世界の穀物生産量は、1965 年の 1006 百万トンから1988 年には 1743 百万トンと世界全体で 73%増加し、特に、開発途上国では同期間に 106%の大きな伸びを記録した。</s>	<div4 type=subsection id="E2.1.1.4"> <head id="E2.1.1.4-h"> (4) Expansion in the production and consumption of agricultural, forest, and marine products </head> <p id="E2.1.1.4-1"> <s id="E2.1.1.4-1.1">By supplying food and timber products, the agriculture, forest, and marine products industries provide the most basic support for human existence.</s> <s id="E2.1.1.4-1.2">In regions involved in agriculture, forest, and marine product related activities, the environmentally-conscious pursuit of these activities is helping maintain the environmental-protection capabilities of agricultural regions and forests.</s> <s id="E2.1.1.4-1.3">On the other hand, as production activities are carried out, a load is placed on the environment because of changes in the intended use of resources.</s> <s id="E2.1.1.4-1.4">Examples of this include the conversion of forests to agricultural land, primarily in developing countries, the fouling of water from the use of fertilizers in Europe and North America, and the emission of methane gas (<cf><mol><a>C<a>H<n>4</mol></cf>), a type of greenhouse gas, by livestock.</s> <div5 type=subsubsection id="E2.1.1.4.1"> <head id="E2.1.1.4.1-h"> A. The production of staple foods</head> <p id="E2.1.1.4.1-1"> <s id="E2.1.1.4.1-1.1">World grain production rose from 1,006 million tons in 1965 to 1, 743 million tons in 1988, a 73% increase.</s> <s id="E2.1.1.4.1-1.2">In developing countries during this same period, there was a sharp 106% rise in grain production.</s>

Figure 1: Examples of the corpus

3. Computerization and SGML Tagging

Some of the white papers are available on CD-ROMs or floppy disks and others are available only in a printed form. The latter, we had to input either manually or by using an OCR. We are formatting our corpus in TEI format using the following steps:

(1) Definition of document type.
We define the document type of our bilingual corpus based on the TEI Lite regulation and its extensions. For chemical formulas, we adopted STANDCOM.DTD in ISO/IEC TR 9573-11. (Burnard, 1995; Bonhomme et al, 1995; Maler and Andaloussi, 1996)

(2) Conversion of nonstandard characters.
Gaiji (nonstandard characters) in Japanese, are converted into some combinations of standard characters. For example, "1 in a circle" is converted into "&c-1;".

(3) Regularization of titles and bodies.
Before tagging bilingual texts, we have to regularize the texts so that we can identify their titles and bodies automatically. We did this regularization process manually because the titles in the English versions tend to be very different from the titles in the Japanese versions.

(4) SGML tagging.
After the regularization, most of the tagging, e.g., (a) identification of the hierarchy of sentences, (b) identification of titles, (c) identification of paragraphs, and (d) identification of sentences, can be done automatically. We are using only part of the tags defined by TEI Lite, e.g., tei, teiHeader, text, body, div, head, p, s, and q. Tasks which we have to do manually, e.g., assigning alignment attributes and identification of quotations, still remain to be done.

As for the character code, this bilingual corpus utilizes JIS (Japanese industrial standard) X 201 and JIS X 0208 for the Japanese text and JIS X 201 for the English text. They can be easily converted into EUC code.

For Entity Sets, we utilize public entity sets such as ISOlat1, ISOgrk3, ISOpub, ISOnum, and ISOamsr. We have, also, defined our own entity set.

4. Sentence to Sentence Alignment

In the 1997 fiscal year, we aligned Japanese sentences with English sentences. Alignment data is a set of one-sentence to one-sentence correspondences from two sets of sentences extracted from corpora in Japanese and English. When one Japanese sentence (J1) is translated into several English sentences (E1, E2, ..., Ek), the correspondence is represented as {(J1, E1), (J1, E2), ..., (J1, Ek)}. Aligned data is developed via automatic processing by using an alignment tagger and by manual post-editing.

Automatic processing is done by software developed by NTT Communications Laboratory (Haruno and Yamazaki, 1997). This software is an accurate and robust text alignment system for English and Japanese. This system utilizes a bilingual dictionary of general use and the word correspondences that are statistically acquired in the alignment process, to avoid a limitation on the amount of word correspondences that can be statistically acquired. This limitation is mainly caused by the big differences in the systems of functional words between Japanese and English. This system, by combining two kinds of word correspondences, gradually determines sentence pairs that correspond to each other by relaxing parameters and align texts of various length with high precision.

Tools for the post-editing of bilingual data and for data conversion have been developed by the committee. The post-editing tool has a graphical user interface to make the process of post-editing efficient. It is written in Tcl/TK and run on UNIX or Windows 95/NT with Japanese Tcl/Tk. In our experience, post-editing takes one minute per sentence. The input to the alignment process is an SGML tagged parallel corpus as described in the above section. The conversion tool removes unnecessary tags from the corpus and converts tags for sentence and paragraph delimiters into a format suitable for the automatic alignment tagger. The tagger analyzes the input and generates alignment data automatically. Sentences are divided differently in the SGML tagged corpus and the automatically aligned corpus. Therefore, a conversion is done to adjust them.

Next, using the post-editing tool, automatically-aligned data is post-edited manually, to correct correspondences between sentences in Japanese and English. This tool is activated by a data pair, i.e., Japanese text and English text. Figure 2 shows a display of the post-editing tool. In the display, Japanese text is shown on the left side and English text on the right. Correspondence relations are represented via lines between them. Post-Editing is done by adding lines, deleting lines and changing attributes of lines. There are three attributes, i.e., automatically aligned, manually added and manually removed. These attributes are represented by different line colors. The post-editor can add comments on the information obtained during post-editing. Data about links and comment data are saved into files when the post-editing is complete.

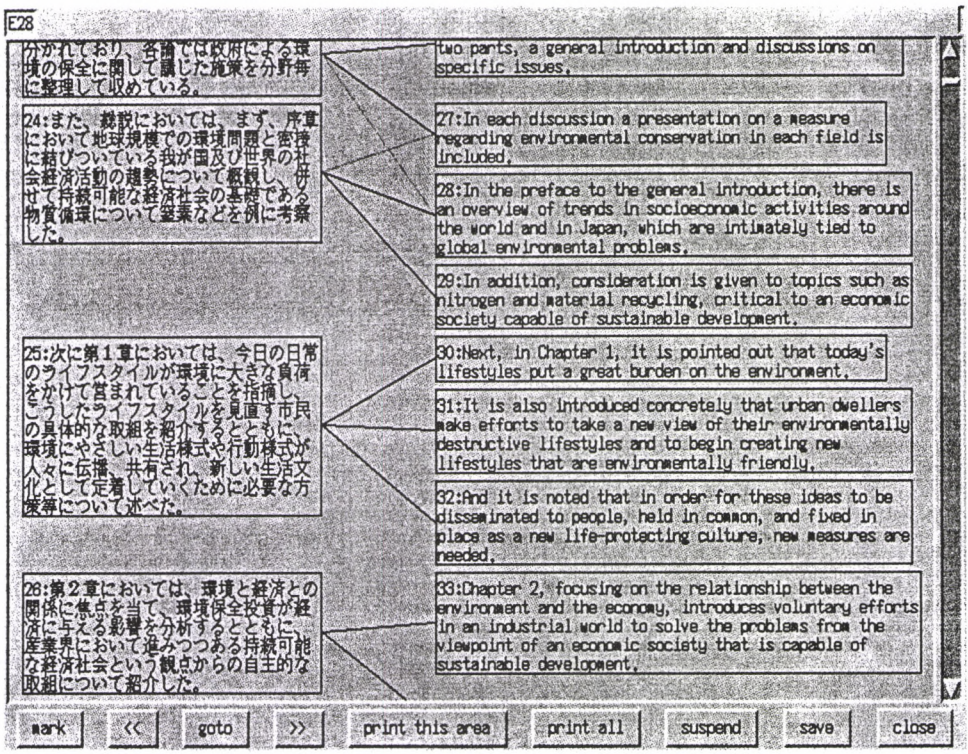


Figure 2: The post-editing tool for a bilingual aligned corpus

Finally, post-edited aligned data is converted into an SGML format. Correspondence information is tagged "TRANSLATION". This has two possible attributes, i.e., "FROM" and "TO". For example, when (the whole or part of) the Japanese sentence "J2.1.1.4-1.1" is translated into (the whole or part of) the English sentence "E2.1.1.4-1.1", the relationship is represented as <TRANSLATION FROM="J2.1.1.4-1.1" TO="E2.1.1.4-1.1">. The aligned data is a set of these data.

The sentences in Figure 2 are from the white paper of the Environment Agency for the 1995 fiscal year. Examples of English translations from Japanese "noun + NO" structures in this white paper are shown in Table 2. "NO" is a Japanese postpositional which is supposed to be similar to the English preposition "of". However, Table 2 shows that such direct translation from "noun + NO" into "of + noun" is rare and various styles of translations occur in the document. Gathering these examples with contexts, we can extract translation rules and make machine translation of higher quality.

Table 2: Examples of English Translations from Japanese "noun +NO" structure

今回の (konkai NO)	this year's	use of " 's "
政府の (seifu NO)	(White Paper)	implicit
初の (hatsu NO)	first	adjective
これまでの (koremade NO)	up to now	idiomatic expression (adverbial)
6月の (rokugatsu NO)	in June	use of " in "
6月の (rokugatu NO)	-	none
今回の (konkai NO)	this year's	use of " 's "
第26回目の (dai 26 kaime NO)	White Paper , the 26th	apposition
基本法の (kihonhou NO)	of the Basic Environment Law	use of " of "

5. Phrase and Proper Noun Level Alignment

The first version of our corpus has only SGML tagged text, sentence-level alignment data and manually added comments. We are trying to add more tags to our bilingual corpus in order to get more precise information, such as part-of-speech tags, word alignment tags, clause alignment tags and syntactic and semantic tags. Currently, we are adding phrase level alignment tags and correspondences between proper nouns and compound nouns in Japanese and English. We have developed assistance tools for these processes as the extension of the post-editing tool described above. This tool has a special window for phrase level and proper noun level tags. The window has three text boxes and several buttons as shown in Figure 3. The left box contains Japanese sentences, the right box contains English sentences, and the center box contains the correspondence between phrases and proper nouns.

Phrase alignment tags are useful since phrase level correspondences are important in MT research. More general linguistic information can be extracted from these than from sentence alignment tags. This can be used as fundamental data for example based machine translation systems. In the first version, correspondences between proper nouns were written manually by the post-editor, in a comment. Those are now stored in a fixed format in the corpus, to be used, for example, as test data for an information retrieval system.

Tagging is done based on the Japanese text, and the correspondence in English is assigned. Firstly, phrase level alignment and proper noun alignment are done, then their correction and compound noun alignment are done. In phrase level alignment, not only direct correspondences but also non-direct correspondences caused by the syntactic differences between English and Japanese are marked. Such correspondences include verb-noun conversion between the two languages and conversions such as "my

country (in Japanese)" and "Japan (in English)" and "same year (in Japanese)" and "1998 (in English)". For proper noun level alignment, words stored in the ordinary dictionaries do not have to be tagged. However, a word which is a proper noun in one language and a common noun in the other language is tagged.

Since the automatic alignment tagger generates correspondence data during the tagging process, we can use this information as a hint for the post-editor. Using the correct correspondence of proper nouns post-edited by hand, the automatic tagger would be able to re-tag the original parallel text more precisely, and this correspondence could be used to improve the quality of the automatic tagger.

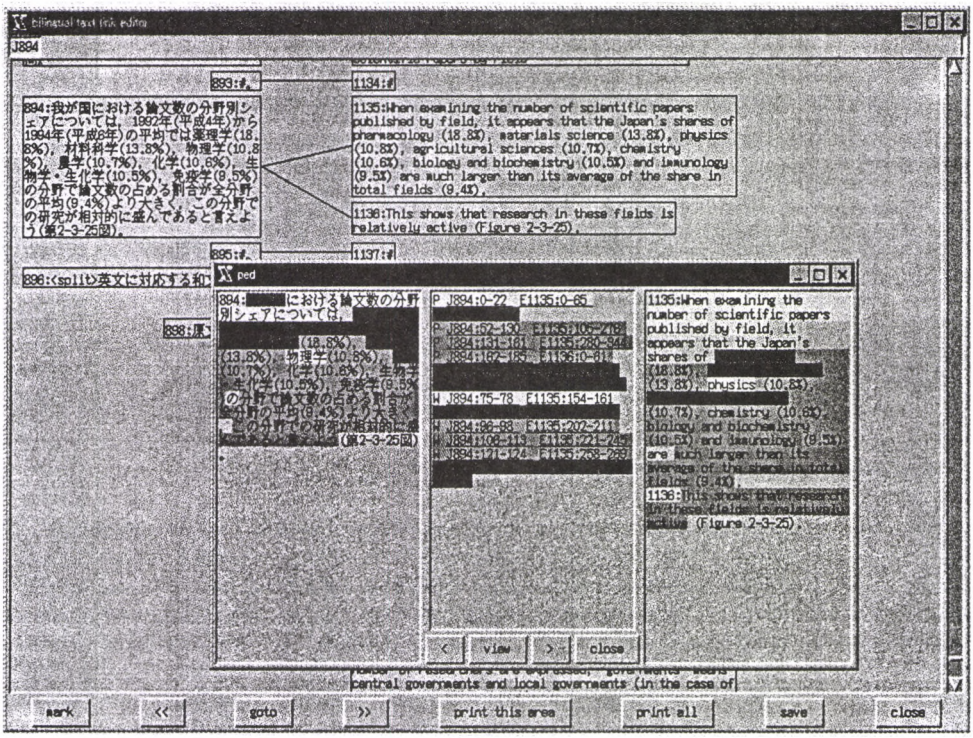


Figure 3: The post-editing tool for a bilingual aligned corpus

6. Conclusion

In this paper, we have discussed JEIDA's bilingual corpus project. This corpus is being developed to be:

(1) available without charge to the public for research and evaluation of NLP technology,
(2) built under cooperation and dispersion of tasks with other research organizations,
(3) general and independent of any one specific linguistic theory.

We will continue our efforts to enlarge this public bilingual aligned corpus for NLP research on these principles. We aim to develop a corpus ten times bigger than the one that we have now.

References

Bonhomme P. et al.(1995). LINGUA Information & Technical Aspect. Lingua Project.

Burnard, L. (1995). TEI Lite: An Introduction to Text Encoding for Interchange. C. M. Sperberg-McQueen.

Haruno, M. and T. Yamazaki (1997). High-performance bilingual text alignment using statistical and dictionary information, Natural Language Engineering, Vol. 3, No. 1., Cambridge University Press.

Maler E. and J. El Andaloussi (1996). Developing SGML DTDs From Text to Model to Markup, Prentice Hall PTR.

Extracting Semantic Similarities of Japanese Adnominal Constituents from Large Corpora

KYOKO KANZAKI – HITOSHI ISAHARA

We were treating adjectives, nominal adjectivals, adnouns and "noun + NO" structures as Japanese adnominal constituents. Among them, only the "noun +NO" structure is a complex phrasal structure. "Noun + NO" structures can represent a wide range of semantic relations, such as adnominal usage, apposition and possession, therefore, their semantic behavior is sometimes similar to and in other time different from the behavior of other adnominal words. Our aim is to establish a basis for this criteria using very large corpora in Japanese. In this paper, first we discuss the classification of the usage of Japanese adnominal constituents briefly, and explain how we extract occurrences of "noun + NO" structures and related adnominal words from the corpora. As a basic idea of this analysis we focus on the phenomena where adnominal constituents represent the concrete contents of their head noun. This makes it possible to identify adjectives and "noun + NO" structures which are similar in semantic behavior to the referents of their head nouns. These expressions are extracted semi-automatically from large corpora. We describe what we learned about the similarities between adnominal words and "noun + NO" structures extracted from the corpora. In this paper we explain it in regard to spiritual activities, <emotion>.

1. Introduction

Our recent research has consisted in classifying relations between adnominal constituents and their modified nouns based on their behavior in actual sentences. We were aiming at handling their meanings dynamically. Such relationships include not only direct "feature-value" relationships in which adnominal constituents are filled as a value of the features of modified nouns, but also the relations which do not come directly from the features of modified nouns. This includes relations connected by contextual information and the case where the adnominal constituents represent the state of being of the referent of the modified nouns (Kanzaki 1997).

We were treating adjectives, nominal adjectivals, adnouns and "noun + NO" structures as adnominal constituents. Among them, only the "noun + NO" structure is a complex phrasal structure. "noun + NO" structures can represent a wide range of semantic relations, such as adnominal usage, apposition and possession, therefore, their semantic behavior is sometimes similar to and in other time different from the behavior of other adnominal words. "NO" is the Japanese postpositional whose meaning is similar to "of " in English.

In our previous research, we treated only the adnominal usage of " noun + NO" structures. Therefore, it was not necessary to establish a criteria to decide whether the usage was adnominal or not. However, when we compile a lexicon for natural language understanding, which treats the meanings represented by "noun + NO" structures differently according to their usages, there must be a criteria to classify the differences in these structures found in actual sentences.

Our aim is to establish a basis for this criteria using very large corpora in Japanese. In this paper, first, we discuss the classification of the usage of Japanese adnominal constituents briefly, and explain how we extract occurrences of "noun + NO" structures and related adnominal words from the corpora. We describe what we learned about the similarities between adnominal words and "noun + NO" structures extracted from the corpora.

2. The Diversity of Semantic Relations between "noun + NO" and their Head Nouns

Among Japanese adnominal constituents, "noun + NO" represents a wider range of semantic relations than other adnominal constituents, therefore, "noun + NO" does not always behave like the other adnominal constituents.

In previous work, some researchers have analyzed semantic relations between the noun in the "noun + NO" structure and its head noun (Shimazu et al.1986). Here, we show several examples that demonstrate the diversity of the semantic relation between "noun + NO" and their head nouns according to their research.

DENWA NO SECCHI	setting of the telephone
DENSHA NO TUUKIN	commuting by train
ASHITA NO DEITO	dating tomorrow
BILU NO MAE	in front of the building
KODOMO NO NAMAE	the name of the child
BAKUHATSU NO GEN'IN	the cause of the explanation
KAISHI NO JIKOKU	the time of beginning
HEYA NO BANGOU	the number of the room

KANOJO NO NOUTO	her note
BENGOSHI NO SMITH SAN	Mr. Smith who is a lawyer

These semantic relations between "noun + NO" and their head nouns are different from those between other adnominal constituents, e.g., adjectives, and their head nouns. However, some "noun + NO" behavior is similar to the behavior of adjectives and nominal adjectivals. In these cases "noun + NO" seems not to differ semantically from adjectives and nominal adjectivals.
Let us consider the English examples.

financial world /world of finance ("ZAIKAI")
industrial center /center of industry ("SANGYOU NO CHUUSHIN")

In this case "noun + NO" need not be distinguished from adjectives with respect to semantic behavior. However, in the following examples it is necessary to distinguish them from each other.

global center /center of globe ("SEKAI NO CHUUSHIN/ CHIKYUU NO CHUUSHIN")

We do not have a discrimination criteria to recognize whether a "noun + NO" structure is similar in its semantic behavior to that of adjectives or not. We have attempted to gather nouns in the "noun + NO" structure which behave like adjectives.

3. Classification of the Usage of Japanese Adnominal Constituents

Before describing the method used in this research, it is necessary to explain the concept which forms the basis of our analysis. We have focused on the semantic relations between adnominal constituents and their head nouns. In this section we briefly mention our previous research.

3.1. Three patterns of syntactic paraphrases of adnominal constituents in attributive position

On consideration of the syntactic relations between adnominal constituents and their head nouns, we find that some adnominal constituents can appear both in the attributive and predicative positions (Sakuma 1967, Martin 1975, Makino & Tsutsui 1986). However, some adjectives express different meanings when they appear in one or the other position and some adjectives can appear only in one of these two positions (Hashimoto & Aoyama 1992).

(A) A paraphrase can be made without changing the modifying relations semantically.

Ad. + noun –> noun GA Ad. (noun "is" Ad.)

(Ad. = adnominal constituent, here an adjective)

(B) A paraphrase can be made, only when a noun is restricted by its context: the presence of modifiers or determiners, e.g. articles.

Ad. + noun –> SONO noun WA Ad. ("that" noun "is" Ad.)

(C) A paraphrase cannot be made at all, i.e. only the attributive position is available.

Ad. + noun –> *NONE*

We can classify semantic relations between adnominal constituents and their head nouns into three types by the use of paraphrases. Paraphrases exist for both Type A and Type B. However, paraphrases cannot be made at all for Type C. This difference is based on the fact that adnominal constituents in types A and B modify the referents of their modified nouns while adnominal constituents in Type C do not modify their head nouns directly.

3.2. Classifications of the Semantic Relation between Adnominal Constituents and their Head Nouns

It is important for the analysis of adjectives to consider what its head nouns denote in the Sentences (Bouillon 1996). Also, when we analyze the word meanings, it is important to take context and our world knowledge into account (Pustejovsky 1995, Lascardies & Copestake 1998). In this section, we briefly describe the semantic relation between adnominal constituents and their head nouns. In a previous paper, we analyzed these semantic relationships in detail (Kanzaki & Isahara 1998) and we mentioned the formal treatment of these relations for an NLP system (Isahara & Kanzaki 1999).

Adnominal constituents modify nominals syntactically and most of these modify their head nouns semantically. Here, the "analysis" of the relationship between adnominal constituents and their head nouns consists in the selection of the attribute of the modified nouns to which the adnominal constituents add some information. In type A, the choice of the attribute can be predicted from the meaning of the referent of their head nouns. In type B, we must identify the attribute of the head noun which is filled by the adnominal constituents. This identification can be made using word meanings directly, inferences from the information in the lexicon, or context.

As an example of Type A, there is "YURUYAKA_NA (gentle) KEISHA (slope)." The adnominal constituent "YURUYAKAN_NA (gentle)" expresses an attribute of an instance of the concept "KEISHA (slope)." The instance "KEISHA (slope)" predicts "an angle of steepness" whose value is either a number or a measure of intensity. This prediction is obtained from the meaning of the modified noun "KEISHA."

As an examples of Type B, there is "OOGARA_NA (large) OTOKO (man)." For example, "man" has several major attributes, e.g. name, age, character, and physique. "OOGARA_NA (large)" can appear in predicative position, i.e., "SONO OTOKO WA OOGARA_DA (that man is large)," with the same meaning that the man's physique is large. We determine the attributes of the referent of the head noun from our world knowledge and then we select the attribute of the modified noun that adnominal constituents like "large" can embody. In this case, "large" describes "physique" an attribute of "man."

However, in Type C, the way that adnominal constituents modify their head nouns is different from Type A and Type B. In Type C adnominal constituents modify (1) only a part of the meanings which their modified nouns allow, (2) the contents of the referents of their modified nouns, or (3) the states of being of the referents of

their modified nouns. Among these three modification relations in Type C, we focus on the second relation, that is the behavior of adnominal constituents which represent the concrete contents of the referent of their head nouns.

4. To explore the Similarities of Semantic Functions between "noun + NO" Structures and Adjectives. (The Method for this Research)

4.1. The Basic Concept

There is one case where the paraphrasing discussed in section 2 is not possible. This occurs when the meanings of adnominal constituents are semantically similar to the features of the referents of their head nouns, e.g., adnominal constituents represent the concrete contents of their head nouns. Let us consider the Japanese phrase "KANASHII KIMOCHI (sad feeling)" and "YOROKOBI NO KIMOCHI (feeling of delight)" as examples.

KANASHII	KIMOCHI
{adjective}	{noun}
(sad)	(feeling)

YOROKOBI	NO	KIMOCHI
{noun}	{postp.}	{noun}
(delight)	(of)	(feeling)

The adjective "KANASHII (sad)" cannot appear in the predicative position without changing the meaning of the phrase. Here, "KANASHII (sad)" can be considered as denoting the concrete content of "KIMOCHI (feeling)." Therefore, both "KANASHII (sad)" and "KIMOCHI (feeling)" include the same semantic element, <emotion>. "YOROKOBI NO KIMOCHI (feeling of delight)" also contains the same semantic relation. "YOROKOBI NO (delight)" represents the concrete contents of its head noun "KIMOCHI (feeling)," therefore, "YOROKOBI NO (delight)" and "KIMOCHI (feeling)" include the same semantic element, <emotion>.

As is seen above, both "KANASHII (sad)" and "YOROKOBI NO (delight)" represent the content of the referent of the head noun whose semantic element is <emotion>, therefore, "KANASHII (sad)" and "YOROKOBI NO (delight)" must be classified into the same semantic category, <emotion>, even though they are classified as different syntactic categories. In these examples, both the adjective and "noun + NO" are classified in the same semantic category, <emotion>.

However, if adnominal constituents do not include the same concept as their modified noun, they can not represent the content of the referent of the head noun. In the following examples, the noun in "noun + NO," "JOHN," does not include the concept, <emotion>, it can not represent the content of "KIMOCHI (feeling)." The adjective, "KANASHII (sad)," and the noun in the "noun + NO," "JOHN" do not embody the same concept and have a different semantic relation with their head noun. We cannot find the semantic similarities between "KANASHII (sad)" and "JOHN" as we could between "YOROKOBI" and "KANASHII."

KANASHII	KIMOCHI
{adjective}	{noun}
(sad)	(feeling)

JOHN	NO	KIMOCHI
{noun}	{postp.}	{noun}
(John's)		(feeling)

We focus on the phenomena where adnominal constituents represent the concrete contents of their head nouns. This makes it possible to identify adjectives and "noun + NO" structures which are similar in semantic behavior to the referents of their head nouns. These expressions are extracted semi-automatically from large corpora.

4.2. How to Extract the Necessary Information

When we collect words which have some similarities, it is difficult to select the semantic axis for classification by making use of only the co-occurring words. In collecting similar words, some previous research took not only co-occurring words but also the context of these words into account (Grefenstette 1994). One of the important points of our analysis is the introduction of the distinct semantic elements that both "noun + NO" structures and adjectivals (adjectives and nominals) have in common with their head nouns. We want to ascertain the similarities between "noun + NO" and other adnominal constituents based on these common semantic elements. For this reason, we use the semantic relations, in which adnominal constituents represent the concrete content of their head nouns, as a key to classification. We extracted these relations from one year of newspaper articles from Mainichi Shimbun (1994), 100 novels from Shincho publishers and 100 books covering a variety of topics. We used the following procedure to extract the necessary information.

Step 1) Extract from the corpora, all nouns which are preceded by the Japanese expression "TOIU" which is something like "that" or "of." "TOIU + noun (noun that/of ...)" is a typical Japanese expression which introduces some information about the referent of the noun, such as apposition. Therefore, nouns found in this pattern may have their content elucidated by means of their modifiers.

Step 2) Extract from the corpora, all "noun + NO" structures, adjectives and nominal adjectivals which modify the nouns extracted in step 1.

NB, the relationships between adnominal constituents and their modified nouns extracted here include not only representations of the contents of the noun, but also other various relations.

Step 3) Extract "noun + NO" structures, adjectives and nominal adjectivals which represent the contents of the referents of the modified nouns.

Step 3 is done by hand. However, to extract these adnominal constituents efficiently, we sorted the words extracted in step 2 using the Japanese thesaurus "Word List by Semantic principles (Bunrui Goi Hyo, NLRI 1964)."

Step 4) In order to find the distribution of their semantic categories and analyze the semantic similarities between "noun + NO" and other adnominal constituents in each semantic category, we clustered the modified nouns automatically. This clustering was based on sets of similar adnominal constituents which represent the content of the referent of the modified noun.

5. Clustering of the Modified Noun

In this section, we show the result of clustering the modified nouns. The modified nouns are clustered by the frequency of coinciding adnominal constituents which appear in each group.

Figure 1: The result of clustering of the modified noun

<1> KANTEN KENCHI BUNYA MEN
(viewpoint standpoint field aspect) → rough meanings
<2> IKI HAN'I AIDA NAKA
(limits sphere among in)
<3> KUUKI KEIKOU
(atmosphere tendency)
<4> INSHOU KIHUU TOKORO TATAZUMAI
(impression character point appearance)
<5> KI OMOI KIMOCHI NEN KANJOU KANJI KAN
(intention thought mind inclination emotion feeling sensation)
KIBUN JO
(mood passion)
<6> IKEN KOE SHISEI IKOU
(opinion calls course intention)

.....and so on.

For example, "KANASHII (sad)," "URESHII (pleasure)," "ZANNEN_NA (regret)," "KANASHIMI NO (of sadness)," "YOROKOBI NO (of delight)" and so on, modify nouns such as "KI (intention)," "OMOI (thought)," "KANJI (emotion)." Using a set of adnominal constituents, such as "KANASHII (sad)," "URESHII (pleasure)," "ZANNEN_NA (regret)," as keys for classification, we can classify modified nouns, "KI (intention)," "OMOI (thought)," "KANJI (feeling)" and so on, into the same group. Then we can find a semantic relationship between these adnominal constituents and their head nouns, in this case, <emotion>. In the following section we describe the similar and different semantic behaviors between "noun + NO" and other adnominal constituents in the same semantic category according to the result of the clustering.

6. Adjectives and "Noun + NO" Structures which Show Similar Semantic Behavior.

As we described in the previous section, we extract sets of "noun + NO" structures and adjectives from the data which was sorted semantically. Words in each set represent the semantic substance of the similar nouns which they modify. Therefore, their semantic categories are similar. We group up these adnominal constituents and their modified nouns, and assign semantic categories for these groups. "Noun + NO" structures and adjectives in each semantic category represent some semantic similarity.

The following example treats modified nouns related to emotion. Examples of modified nouns of a similar semantic category and their modifiers which have a semantic category similar to that of the nouns are listed below. Included are some "noun + NO" examples which though co-occurring with <emotion> nouns are not classified as such themselves.

Figure 2: The modified nouns and adjectives, nominal adjectivals, and "noun + NO" collected in the semantic categories, <emotion>

Modified nouns:
KANJI (feeling), KAN (sensation), OMOI (thought), KI (intention), NEN (inclination), KIMOCHI (mind), KIBUN (mood), KANJO (emotion), JO (passion)
Adjectives and nominal adjectivals:
AWARE_NA (poor), IJIRASHII (moving), HOKORASHII (triumphant), KINODOKU_NA (unfortunate), SHIAWASE_NA (happy), ZANNEN_NA (disappointing), URESHII (pleasure) ...and so on.
"Nouns" in "noun + NO" structure
a) spiritual activity KANASHIMI (sadness), FUKAI (displeasure), SHITASHIMI (familiarity), ZOUO (abhorrence), GAMAN (endurance), KOUKAI (regret), YOROKOBI (joy), MANZOKU (satisfaction), RAKUTAN (disappointment), IGAI (unexpected), ...and so on. **b) mainly action nouns** HOSHIN (self-defense), CHIKUZAI (moneymaking), INTAI (retirement), HIHAN (criticism), HIYAKU (rapid progress), HEIWA (peace) ...and so on

There are many adjectives and nominal adjectivals which can modify nouns in Figure 2, such as "AWARENA (poor)," "IJIRASHII (moving)" and "HOKORASHII (triumphant)." Some "noun + NO" structures are semantically similar to these adjectives since they represent the contents of the emotion, e.g., "FUKAI NO KAN (displeasure feeling)" and "YOROKOBI NO KIMOCHI (emotion of delight)." Most nouns in these "noun + NO" structures in Figure 2 are classified into "spiritual activity by humans (spirit)" by the "Word List by Semantic Principles." "Noun + NO" structures which have this kind of semantic category are similar to adjectives and nominal adjectivals, as both represent the content of the human mind, i.e., thought. We call this semantic category created by these adnominal constituents and their modified nouns "Feeling."

On the other hand, some adnominal relationships concerning emotion can only be represented by "noun + NO" structures, such as "HOSHIN NO KIMOCHI (desire of defending one's own interest)," "CHIKUZAI NO NEN (thought of moneymaking)" and "INTAI NO KIMOCHI (thought of retirement)." Action nouns are mainly used in these "noun + NO" structures. Adnominal modifying relations of "action_noun + NO + emotion_noun" structures represent positive or intentional emotion. This kind of intentional emotion cannot be

expressed by adjectives. We call this semantic category "Intentional Emotion.."

We discuss two types of semantic representations above, i.e., Feeling and Intentional Emotion. Feeling can be represented by adjectives and "noun + NO" structures. However, Intentional Emotion can be represented only by "noun + NO" structures. From the standpoint of the characteristics of the modified nouns (they represent human emotions), these two spiritual activities (Feeling and Intentional Emotion) are similar, even though there are differences in whether the activity is intentional or not. However, from the standpoint of the selection of adnominal relationship in the surface structure, whether the activity has active intention or not will be the important factor for the selection between adjectives and "noun + NO" structures.

There is an exception to the above that only "noun + NO" structures can be used with intentional emotion. In the case of the Japanese noun "ISHI" whose meaning is an intentional emotion, e.g., "the will," other adnominal words may be used, including certain adjectives. Let us compare them at the following list.

Figure 3: List of adnominal constituents that co-occur with "ISHI (intentional emotion)"

Adjectives and nominal adverbials
TSUYOI (strong), KATAI (firm), MEIKAKU_NA (clear), SHIZEN_NA (spontaneous), JIYUU_NA (free)...and so on
Nouns in "noun + NO"
KIN'EN (quitting smoking), KAIKO (opening a sea port), KIKOKU (returning home), KOUHUKU(surrender), SHUUSHOKU (finding a job) ...and so on

Action nouns occurring in these "noun + NO" structures represent the action with positive or intentional emotions when these nouns co-occur with "ISHI (intentional emotion)." On the other hand, some adjectives, such as "TSUYOI (strong)" and "KATAI (firm)," also modify "ISHI (intention)." These adjectives describe the quality of the intention, e.g., the strength of intention. Here, both "noun + NO" structures and adjectives can modify "ISHI (intentional emotion)," however, there is a difference in the semantic function between adjectives and "noun + NO" structures.

7. Conclusion and Future Trends of the Research

Sometimes, it is necessary to distinguish the semantic modification functions between adjectives and nominal adjectivals in adnominal usage and "noun + NO" structures as adnominal constituents. At other times, this is not necessary. However, there is no explicit criteria to discriminate between these two possibilities. In this work, to find the key for classification, we collected, from large corpora, "noun + NO" structures which have similar semantic behavior to that of adjectives.

In order to analyze linguistic phenomena empirically, we need to do research using large corpora which reflect the real world surrounding us. It is important to find the best methods for extracting from these corpora, precisely the linguistic phenomena we desire to study. In the future, in addition to studying the methodology for extracting similar adnominal words i.e., "noun + NO" and other adnominal constituents, from corpora, we will continue to try establishing criterion to distinguish the ways "noun + NO" and other adnominal constituents are similar and the ways they differ in their semantic behavior relative to their head nouns.

References

[Kanzaki & Isahara,1998] K.Kanzaki and H. Isahara, The Semantic Connection between Adnominal and Adverbial Usage of Japanese Adnominal Constituents, In Proc. of 10th European Summer School in Login, Language and Information "Lexical Semantics in Context: Corpus, Inference and Discourse" workshop

[Shimazu et al 1986] A.Shimazu, S.Naito and H.Nomura, Joshi "NO" ga musubu meisi_no imikankei no kaiseki, "keiryou kokugogaku" 15-7

[Sakuma 1967] K.Sakuma, Nihonteki Hyougen no Gengo Kagaku, Kousei-sha Kousei-kaku

[Martin 1975] S.Martin, A Reference Grammar of Japanese, Yale University Press

[Makino & Tsutsui,1986] S.Makino and M.tusui, A Dictionary of Basic Japanese Grammar, The Japan Times

[Hashimoto & Aoyama 1992] M.Hashimoto and F.Aoyama, Three usages of adjectives, mathematical Linguistics, 18(5)

[Bouillon 1996] P.Buillon, Mental state adjectives, the perspective of generative lexicon, Proceedings of the 16th International Conference on Computational Linguistics (COLING `96)

[Lascarides & Copestake 1998] A.Lascarides and A.Copestake, Pragmatics and word meaning, Journal of Linguistics, 34(2)

[Pustejovsky 1995] J.Pustejovsky, The Generative Lexicon, The MIT press

[Isahara & Kanzaki 1999] H.Isahara and K.Kanzaki, Lexical semantics to disambiguate polysemous phenomena of Japanese adnominal constituents, Proceedings of the 37st Annual Meeting of the Association for Computational Linguistics (ACL)

[Grefenstette 1994] G.Grefenstette, Corpus-Derived First, Second and Third-Order Word Affinities, Proceedings of the EURALEX `94

NLRI (National Language Research Institute), Word List by Semantic Principles, Shuei Shuppan (1964)

System of Computerized Czech Word-formation

JANA KLÍMOVA

ABSTRACT

The project concerns the development of an algorithmic system for computational processing of Czech word-formation. Such a system is of particular value for a highly inflected language such as Czech.

The aim of my work is twofold: to generate a new word from a given set of stems and affixes and to analyse a given word and determine its derivational base and its paradigmatic and semantic properties.

A chosen set of stems and affixes was described, their paradigmatic and semantic aspects were stated. The functional variability (alternation of stems) and possible combinations of stems and affixes was studied in this work. By using all this information the derivation relations in the process of creating new words were defined. The constraints concerning the respective derivation relations were stated and will serve as the base for the programming tool which will be able to fulfil both aims of this project. All information about stems and affixes as the main means of derivational morphology necessary for defining the word-formation processes was stored in a FoxPro database system.

This project is supported by the Research Support Scheme of the OSI/HESP, grant No.: 1087/1997.

I. Aim of the project

The project concerns the development of an algorithmic system for the computational processing of Czech word-formation.

The vocabulary of our language is not a fixed list of words but a growing and developing store. There are several procedures how to make a new word. New names of objects are created from existing words by certain changes of morphological structure.

Morphology is a part of grammar studying morphems, their forms and functions. The principal means of morphology are
(a) flexion - conjugation and declination,
(b) wordformation.

These changes are defined by derivational rules and constraints. This set of rules and constraints serves as a basis for the programming tool which will fulfil both aims of this project:

(1) generation of new words from a given set of stems and affixes

(2) analysis of a given word and determining its derivational base and its paradigmatic and semantic properties.

II. Process of word-formation: its components

The infinite entity of words can be divided into (see table no.1):

1. SIMPLE (basic, unmotivated) words: words which are combinable with certain types of suffixes.
2. COMPOUND (derived, created, motivated words): words which are created/derived from nouns (verbs, adjectives) by given types of suffixes. Derived word could serve as a base for the creation of a new word.

In Table 1 the relations between the different types of suffixes and basic and newly created words are demonstrated.

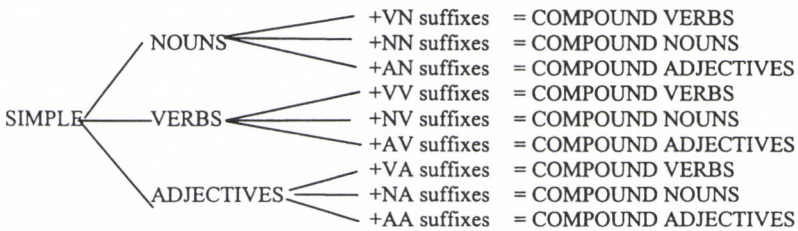


Table 1 Relations between basic and derived words suffixes

The following lists give examples of different types of suffixes and show words derived by these suffixes with their basic words in the brackets.

List of VN suffixes
(serving for the creation of nouns from verbs):

1. -tel, e.g. ředitel (řídit), cestovatel (cestovat), buditel (budit), školitel (školit), pokračovatel (pokračovat), sběratel (sbírat);
2. -č, e.g. řidič (řídit), trubač (troubit), lamač (lámat), sazeč (sázet), hlídač (hlídat);
3. -dlo, e.g. mýdlo (mýt), žrádlo (žrát), chodidlo (chodit), letadlo (létat), prostěradlo (prostírat), kloktadlo (kloktat);
4. -tko, e.g. drbátko (drbat), kukátko (koukat), plivátko (plivat), trsátko (trsat), šoupátko (šoupat);
5. -ák, e.g. šroubovák (šroubovat), piják (pít), věšák (věšet);
6. -ivo, e.g. topivo (topit), hnojivo (hnojit), pojivo (pojit).

List of NN suffixes
(serving for the creation of nouns from nouns):
1. -ář, e.g. kovář (kov), mlynář (mlýn), nástrojář (nástroj), rybář (ryba);
2. -ník, e.g. školník (škola), hrobník (hrob), divadelník (divadlo), hubník (houba), lodník (loď);
3. -ista, e.g. traktorista (traktor), houslista (housle), klavírista (klavír);
4. -ák, e.g. školák (škola), dřevák (dřevo);
5. -árna, e.g. cukrárna (cukr), vinárna (víno), cementárna (cement), kavárna (káva);
6. -írna, e.g. konírna (kůň), brusírna (brus);
7. -ovna, e.g. knihovna (kniha), strojovna (stroj).

List of AN suffixes
(serving for the creation of nouns from adjectives):
1. -ník, e.g. vraník (vraný), jarník (jarní), zimník (zimní);
2. -ost, e.g. hloupost (hloupý), soudnost (soudný), přímost (přímý), velikost (veliký).

List of NA suffixes
(serving for the creation of adjectives from nouns):
1. -ový, e.g. jahodový (jahoda), jablkový (jablko), tiskový (tisk), listový (list);
2. -natý, e.g. dřevnatý (dřevo), kolenatý (koleno), ramenatý (rameno);
3. -ovitý, e.g. houbovitý (houba), šlachovitý (šlacha);
4. -ský, e.g. rajský (ráj), horský (hora), chlapský (chlap);
5. -cký, e.g. otrocký (otrok), světácký (světák), pijácký (piják);
6. -ný, e.g. dřevný (dřevo), senný (seno), vonný (vůně).

List of AA suffixes
(serving for the creation of adjectives from adjectives):
1. -ejší/ější, e.g. světlejší (světlý), dolejší (dolní), hodnější (hodný), křivější (křivý);
2. -avý, e.g. bělavý (bílý), modravý (modrý), zelenavý (zelený);
3. -ičký, e.g. maličký (malý), mladičký (mladý), stařičký (starý);
4. -oučký, e.g. maloučký (malý), hodňoučký (hodný), zlaťoučký (zlatý);
5. -inký, e.g. malinký (malý), prostinký (prostý), lehoulinký (lehký);
6. -ounký, e.g. malounký (malý), hlaďounký (hladký), slaďounký (sladký);
7. -ánský, e.g. velikánský (veliký), hrozitánský (hrozivý), ukrutánský (ukrutný).

List of VA suffixes
(serving for the creation of adjectives from verbs):
1. -ící, e.g. honící (honit), vodící (vodit), pěnící (pěnit);

2. -ivý, e.g. vodivý (vodit), perlivý (perlit), horlivý (horlit).

List of NV suffixes
(serving for the creation of verbs from nouns):
1. -ět, e.g. vonět (vůně), zkrásnět (krása), šumět (šum);
2. -it, e.g. trůnit (trůn), strašit (strach), ztvárnit (tvar), vyrobit (výroba);
3. -ovat, e.g. startovat (start), pochodovat (pochod), válcovat (válec).

List of VV suffixes
(serving for the creation of verbs from verbs):
1. -nout, e.g. vyvinout (vyvíjet), kousnout (kousat), rýpnout (rýpat);
2. -it, e.g. vyrobit (vyrábět), uhodit (házet), zlomit (lámat);
3. -at, e.g. vstávat (vstát), chodívat (chodit), volávat (volat).

List of AV suffixes
(serving for the creation of adjectives from verbs):
1. -ávat, e.g. zmodrávat (modrat).

All necessary information about the principal components of word-formation process was collected. A chosen set of stems and affixes was described, their paradigmatic and semantic aspects were stated.

This data was stored in a FoxPro database system, three databases were created: the database of affixes, the database of stems, the database of semantic codes.

The database of affixes (see Table 2) describes the most productive affixes and gives the basic grammatical information for defining the derivation relations.

Suffix	*POS of derived word*	*Gender*	*Semant*	*POS of basic word*
-an	s	m	OBV	av
-ka	s	mf	P,DIM	sa
-ačka	s	f	DIM	s
-na	s	f	P,M	sv
-árna	s	f	HMO	s
-ina	s	mf	M	as
-ota	s	f	VL	a
-dlo	s	n	M	v

Table 2 Database of affixes

The database of stems describes the most frequent stems.

The database of semantic codes (see Table 3) serves for the semantic description of affixes and stems.

Semantic code	*Explanation*	*Examples of suffixes*
OBV	names of inhabitants	0,-an,-ák,-ec
M	locations	-ina,-dlo,0
P	means	-dlo,-ník
D	acts	-ní,-ba,0

VYS	results of acts	-ek,-ení
HMO	mass	-ina,-ovina
DIM	diminutives	-ek,-ík,-eček,-íček,-ka,-čka,-ko,-íčko
VL	properties	-ina,-ota,-ost

Table 3 Database of semantic codes

Back-ordered dictionaries (dictionaries in which the words are alphabetized starting with the last letter) and corpora are good sources for studying the word behaviour and give material for the information about terminal affixes, about the frequency of their use and about the possible alternations caused by these affixes. I used for my study the lemmatised and tagged Czech National Corpus (CNC) with 130 millions of current words.

By using of all these data the rules for word-derivation (derivation relations and constraints) were stated.

III. Generation and analysis of words

One of the aims of the project is to generate a new word from a given stem and affix.

There are two main limiting factors on productivity in word-formation (Šmilauer 1971, Dokulil 1962):

(a) Each suffix has its function (semantic properties) and gives it to the newly created word. This statement supposes that every affix encodes not only part-of-speech but also semantic function and can be combined only with certain stems from the semantic point of view. The database of semantic codes has to serve as a basis for the semantic description of affixes and stems. This work is not yet finished, it needs lot of manual linguistic work and is significantly user dependent.

In Table 4 I give examples of automatically generated words which are correct from the point of view of alternation but from the semantic point of view they could be

- used (in the following table indicated by „u“) or
- not used (indicated by „n“)

Stem	*Suffix -tel*	*Suffix -č*	*Suffix -dlo*	*Suffix -tko*
nést	nositel(u)	nosič(u)	nosidlo(u)	nosítko(u)
brát	bratel(n)	bráč(u)	bradlo(u)	brátko(n)
mazat	mazatel(n)	mazač(u)	mazadlo(u)	mazátko(n)
péct	pekatel(n)	pekáč(u)	pekadlo(n)	pekátko(n)
hrabat	hrabatel(n)	hrabáč(n)	hrabadlo(u)	hrabátko(u)

Table 4 Automatically generated words

(b) The suffix could be combined only with a certain form of the stem. This statement says that only a certain form of the stem (e.g. the verbal present or past tense stem) could serve as a basic stem for the creation of a new word by given suffix. The database of affixes describes the most productive affixes and gives the basic information for defining the derivation relations not only from the

grammatical point of view but also from the orthographic point of view. The set of words is infinite and still the new ones could be created. These words could be
- correct or (e.g. kniha - knížka)
- incorrect (e.g. kniha - *knihka)

from the point of view of alternation `(i->í, z->ž)`.

It was necessary to find all possible alternations and decide which allomorph is combinable a with given suffix. The derivation constraints were stated as generally as possible. By using these constraints and upon the information about the possible semantic and grammatical combination of stems and affixes the program which will fulfil both aims of this project is being built.

If a word from the text corpus is lemmatised, its paradigmatic and semantic properties could be determined according to its suffix by using the information about affixes, derivation constraints and semantic codes from the set of databases.

IV. Computerisation of the word-formation process

The derivation relations in the process of creating new words were defined by using all the information stored in the database system and by studying the functional variability of stems and possible combinations of stems and affixes in the corpus.

The constraints concerning the respective derivation relations were stated and will serve as the basis for the programming tool which will be able to fulfil both aims of this project.

I concentrated on the class of diminutive suffixes which cause many different alternations and . A large variety of different derivative paradigms has to be defined. Table 5 shows these paradigms and how the type of alternation depends on the part of speech, gender and ending of the basic word. The diminutive suffixes are combinable with most nouns, it means that there are almost no restrictions from the semantic point of view but the rich alternation makes difficulties in the process of computerisation of this type of word-formation.

Rules	Constraints	Examples
masculine gender		
aS - áSek - áSeček		vlas - vlásek - váseček
	ch - š	prach - prášek- prášeček
aj - ajík -ajíček		čaj - čajík - čajíček
ak - áček - áčíček		vlak - vláček
ař - ařík - ařiček		trakař - trakařík - trakaříček
áS - áSek - áSeček		jestřáb - jestřábek - jestřábeček
	h - ž	práh - prážek
ách - ášek - ášeček		hrách - hrášek- hrášeček
áv - ávík - ávíček		páv - pávík - pávíček
áz - azík - azíček		mráz - mrazík - mrazíček
eS - eSík - eSíček		dřep - dřepík - dřepíček
eS - í/ýSek - í/ýSeček	h - z,ch - š	jelen - jelínek - jelíneček
ec - eček		čtverec - čtvereček
ek - eček		písek - píseček
ek - (ík) - íček		klacek - klacík - klacíček
el - ýlek - ýleček		jetel - jetýlek - jetýleček

el - elík - elíček	jetel - jetelíček
el - lík - líček	uhel - uhlík - uhlíček
en - ínek - íneček	jelen - jelínek - jelíneček
eň - eník	učeň - učeník
eň - ínek - íneček	stupeň - stupínek
eň - ýnek - ýneček	oheň - ohýnek - ohýneček
et - tík - tíček	nehet - nehtík - nehtíček
éS - e/éSík - éSíček	chléb - chlebík - chlebíček
ém - émek -émeček	problém - problémek - problémeček
ěS - íSek - íSeček	medvěd - medvídek - medvídeček
ěh - ěžek	výběh - výběžek
iS - íSek - íSeček	hřib - hříbek -hříbeček
íS - íSek - íSeček	díl - dílek - díleček
oS - oSík - oSíček	nos - nosík - nosíček
oS - ůSek - oSeček	hrob - hrůbek - hrobeček
oh - ožík	hloh - hložík
och - ošík	hoch - hošík - hošíček
oj - ojek	stroj - strojek
ok - oček - očíček	blok - bloček
om- omek	strom - stromek - stromeček
uS - ouSek	holub - holoubek
už - užík - užíček	muž - mužík - mužíček
ouS - ouSek - ouSeček	kloub - kloubek - kloubeček
yS - ySek - ySeček	záhyb - záhybek - záhybeček
yj - yjík	pyj - pyjík
yk - ýček	jazyk - jazýček
ýS - ýSek - ýSeček	sýr - sýrek - sýreček
rS - rSík - rSíček	prd - prdík
rch -ršek	arch - aršík
rk - rček	krk - krček
rp - rpek	srp - srpek - srpeček
Sl - Slík - Slíček	štrůdl - štrůdlík - štrůdlíček
Sr - Sřík Sříček	zubr - zubřík - zubříček

Table 5 Derivation rules and constraints for diminutives (S - any consonant)

On the other hand there are some suffixes (e.g. "-tel" or "-ák", equivalent to the English suffix "-er") which cause very few changes in the stem. Most of these words are used very occasionally, it means that it is difficult to decide which word is a correct one from the semantic point of view.

Some nouns derived by the „-tel" suffix found in the Czech National Corpus are shown in Table 6.

-tel noun	*Frequency in CNC*	*Basic verb*
po-zor-ova-tel	4866	pozorovat
pře-krač-ova-tel	1	překračovat
pře-svědč-ova-tel	1	přesvědčovat
pře-vych-ova-tel	1	převychovat
pře-vy-prav-ova-tel	1	převypravovat

při-hlaš-ova-tel	940	přihlašovat
při-prav-ova-tel	8	připravovat
při-stěh-ova-tel	1	přistěhovat
půjč-ova-tel	12	půjčovat
roz-děl-ova-tel	1	rozdělovat
roz-hod-ova-tel	8	rozhodovat
roz-jasň-ova-tel	1	rozjasňovat
roz-množ-ova-tel	3	rozmnožovat
roz-šiř-ova-tel	10	rozšiřovat

Table 6 „-tel" nouns found in the Czech National Corpus

Table 6 shows the morphological structure of derived words. It can be seen that the creation of this type of words is very easy. The most frequent „-tel" noun is „pozorovatel" (observer). Some interesting occasional words (pře-krač-ova-tel, pře-svědč-ova-tel, při-stěh-ova-tel, roz-jasň-ova-tel) could be found in the corpus, too. These words illustrate the individual use of this suffix.

V. Possible applications of this system

This system has both analytic and productive potential. One the one hand, it can be used for the analysis of words in text, associating them automatically with the appropriate basic words and thus enabling a machine-translation program to select the appropriate translation or a linguist or lexicographer to associate each lexical item with the semantic and other interpretation of the underlying lemma. The Czech language is very rich in inflections and affixes. This system is a contribution to the Czech morphological analyser.

From the language production point of view, the system can be used to generate appropriate word forms when encoding the Czech language from an underlying message, e.g. the logical form or from a foreign- language original. It therefore has important machine-translation applications.

In addition to information about individual words, the lexicon will include a set of prototypes of derived lexical items, with definitions of the constraints governing the respective derivation relations; this will enable the lexicon to cover also words (word forms) whose presence in the main lexicon (main list of stems) is unpredictable. Possible applications of this system are various NLP systems, e.g. spelling checkers.

Bibliography:

Slavíčková, E.: Retrográdní morfematický slovník češtiny, Academia 1975

Králík J., Petr J., Těšitelová M.: Retrográdní slovník současné češtiny, Academia 1986

Slovník spisovného jazyka českého 1-4, Academia 1957-1971

Dokulil M.: Tvoření slov v češtině, Academia 1962

Šmilauer V.: Novočeské tvoření slov, SPN 1971

Čermák F. : Syntagmatika a paradigmatika tvoření slov II, Morfologie a tvoření slov, Univerzita Karlova Praha 1990

APOLN: A Partial Parser of Unrestricted Text

ANTONIO MOLINA – FERRAN PLA – LIDIA MORENO – NATIVIDAD PRIETO

Abstract

In this paper, we present APOLN (**A**nalizador **P**arcial de **O**raciones en **L**enguaje **N**atural): a partial parser of unrestricted natural language sentences based on finite-state techniques. Partial parsing has been used in several applications: syntactic parsing of unrestricted texts, data extraction systems, machine translation, solving the attachment ambiguity, speech recognition systems, text summarization, etc. The main attractiveness of partial parsing is that is able to handle unrestricted sentences, that contain lexical errors or that present constructions not accepted by the defined grammar. Partial parsing is an alternative to the definition of wide coverage grammars whose definition is an expensive and complex task and that present well-known problems such as overgeneration, undergeneration and ambiguity. We present APOLN as a tool that can be used to construct syntactically annotated corpora from lexically tagged corpora. We also present the results (precision and recall rates) of applying APOLN on unrestricted Spanish corpora, and how tagging errors influence the performance of the parser.

1 Introduction

One of the current focuses of research within natural language processing is the partial and robust parsing of sentences in natural language. The main attractiveness of partial parsing is that is able to handle unrestricted sentences, that contain lexical errors or that present constructions not accepted by the defined grammar. Partial parsing is an alternative to the definition of wide coverage grammars whose definition is an expensive and complex task and that present well-known problems such as overgeneration, undergeneration and ambiguity. (Partial parsing *aims to recover syntactic information efficiently and reliably from unrestricted text by sacrificing the completeness and depth of analysis* [Abney97]).

Partial parsing uses more robust and more efficient algorithms than global parsing. It works with simpler grammars usually defined with regular patterns. Moreover, it handles mechanisms that allow us to continue the analysis in spite of non-understandable segments of words.

While the output of a global parser is a complete analysis tree, if the sentence is syntactically correct, a partial parser postpones the attachment decisions between grammatical constituents if it does not have enough information. In this case, the output is a forest of subtrees which are not interleaved, that is, the trees do not share any nodes. Each tree represents a parsed fragment of the input. Segments of words that have not been recognized appear between the subtrees.

One of the applications of partial parsing is the syntactic parsing of unrestricted corpora. Partial parsing can be used as a first step to construct a syntactically parsed corpus (with complete analysis trees). Other applications are data extraction systems, machine translation, solving the attachment ambiguity, speech recognition systems, text summarization, etc. An overview of the these applications of partial parsing can be seen in [Molina98].

We report in this paper our experiments with APOLN (**A**nalizador **P**arcial de **O**raciones en **L**enguaje **N**atural), a partial parser of unrestricted natural language sentences based on finite-state techniques. In section 2 we describe the system, the structure of analyser and the features of the tagger used. In section 3 we show the performance of the system through some preliminary experiments on the Spanish corpora Cpirápides and LEXEXP. Finally some conclusions and future work are stated.

2 System Description

APOLN is an incremental parser based on finite-state techniques that allows for syntactic partial parsing of unrestricted text. The system takes as input a tagged sentence and produces its partial interpretation as result of applying a sequence of analysis levels following an architecture similar to the presented in [Abney96], [Aït-Mokhtar97], [Chanod96] or [Ejerhed88]. The goal of each level is to recognize certain syntactic structures which are defined using regular expressions. These levels are organized in cascade, so the output of a certain level is the input of the next level. An outline of the system proposed can be seen in Figure 1.

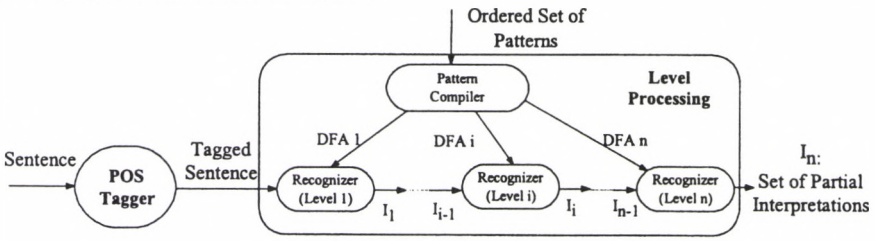


Figure 1: APOLN scheme

2.1 Level Processing Description

A *tagged sentence* and an *ordered set of patterns* which are distributed into a certain number of levels (*n*) forms the input to the *Level Processing* module. These levels are organized in cascade so that the input to a level *i* is the *interpretation* produced by level *i-1* (I_{i-1}). The input to the first level is the *tagged sentence* and the output of the last level (the output of the system) is a *set of partial interpretations* that represents the syntactic parse of the *sentence* using bracketed format.

Each level is defined by a set of patterns using regular expressions. Each set of patterns of a specific level is compiled (Pattern Compiler) into a deterministic finite state automaton (DFA). When the *Recognizer* module is executed for a level *i*, it takes I_{i-1} and the corresponding DFA_i as input. The output (I_i) represents the input I_{i-1} in which the longest sequences of symbols that match a pattern (longest match, [Abney96]) have been identified using boundary markers and syntactic tags. The final state reached determines the matched pattern.

The symbols allowed for defining a pattern of any level are whatever lexical tags and whatever pattern which are defined at a previous level. In this way, patterns are *non-recursive* which allows for incremental parsing. Theses patterns can represent the syntactic constituents, such as noun phrases, adjective phrases, etc., that should be identified from the input sentence.

We have used the usual operators for the definition of the patterns: concatenation, Kleene closure (*), positive closure (+), union (|), one or more cases (?), and parentheses. The set of patterns defined is a set of regular definitions which are grouped by levels. Each level can be defined by several patterns. **Figure 2** shows the scheme for a certain level *i*, where $\mathbf{p_{i,j}}$ is a symbol that represents the *j* syntactic structure expressed at level *i*, and $\mathbf{r_{i,j}}$ is the regular expression that defines the pattern using the indicated operators and the non-recursivity constraint.

Level i //comment
$p_{i,1}$ -> $r_{i,1}$ //comment
$p_{i,2}$ -> $r_{i,2}$ //comment
...
$p_{i,n}$ -> $r_{i,n}$ //comment

Figure 2: Level definition scheme

Note that these patterns can be used to identify specific occurrences such as dates, entities, specific expressions, etc. which could be useful in data extraction systems.

Moreover our approach allows for including actions within the definition of the patterns, by means of *agreement operator* and *inheritance operator*. The agreement operator means that the transition between two states of the DFA is possible when the compatibility condition between two feature structures is true. The inheritance operator indicates the features that have to be inherited from one level to higher levels. This allows for taking into account the morphosyntactic features that are necessary to parse sentences correctly and that can help to solve some parsing errors which are caused by the application of the longest match heuristic. A detailed description of these aspects can be found in [Molina99].

Input and output format is bracketed text, which is similar to the format used for parsing large corpora of text (e.g. Penn Treebank [Marcus93]). The input and the output of each level of processing is composed of a sequence of symbols $s_1\ s_2\ ...\ s_m$, where each s_i can be a **lexical tag**, a **pattern** defined at a previous level, or a boundary mark (beginning, **[**, or ending, **]**, marks). A pattern symbol always appears after an ending mark. So, if $s_1\ s_2\ ...\ s_m$ is an input string, p_i is a pattern of level *i*, given that there exists a sequence of *k* symbols that matches p_i from position *j*, the output would be the sequence $s_1\ s_2\ ...[\ s_j\ s_{j+1}...s_{j+k-1}\ ]p_i...\ s_m$. The output of the last level corresponds to the set of partial interpretations of the input sentences. In **Figura 3** can be seen a sentence which has

been parsed after two levels of processing. In the first one have been identified noun heads (NSN) and verbal head (NSV) and in the second one noun phrases (SN).

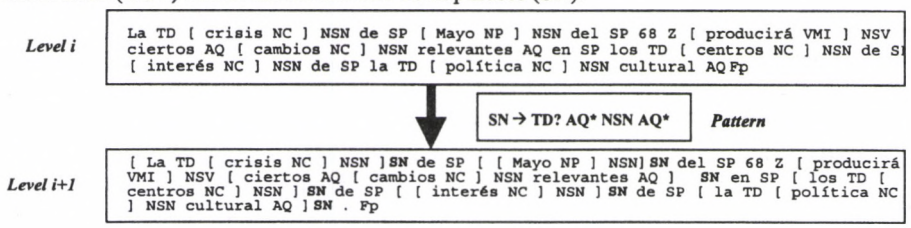


Figure 3: An example of parsed sentence.

2.2 POS Tagger description

A tagger can be considered as a translator that inputs strings from a certain language (Unrestricted Text) and outputs the corresponding sequence of lexical tags or grammatical categories (Text Tagged). Generally, these categories are taken from a set defined previously by linguistic criteria. When a word can be assigned to different lexical categories, the disambiguation is solved by using the information of the context in which this word appears. **Figure 4** shows an scheme of the tagger used.

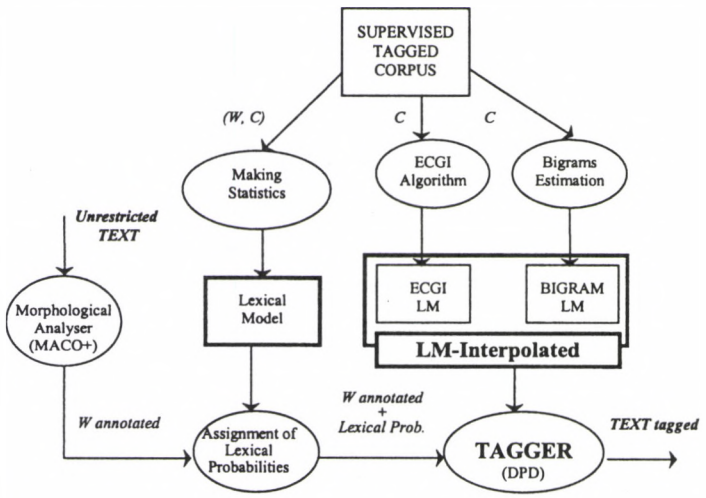


Figure 4: Outline of the POS tagger

The tagging process involves two knowledge sources: the *language model (LM)*, which describes the possible (or probable) sequencing of the categories, and the *lexical model* which represents the relationships between the vocabulary of the application and the set of categories.

The language model used is a stochastic regular grammar or finite-state automaton learnt automatically from data (sequences of categories C) using grammatical inference techniques; in particular, we have used the ECGI algorithm, [Rulot89], [Prieto92]. The model learnt generalizes the sequence of POS strings in the training corpus. In order to increase the coverage of the ECGI model, it has been smoothed by linear interpolation with a simple bigram model. The interpolation

factor has been estimated experimentally. Then, a renormalization process is applied in order to maintain the stochastic consistency. A more detailed description can be seen in [Pla98].

The lexical model has been estimated as usual from a manually tagged corpus (W,C) by computing words, categories and words per category frequencies.

The input of the tagger (Unrestricted Text) is processed by the modular morphological analyser MACO+ [Carmona98]. It performs a proper segmentation of the input text into tokens, identifying punctuation marks, lexical units, dates, abbreviations, numbers, proper nouns, etc. Also, it supplies all the possible lexical tags for every token detected. Then, the lexical probability of each output token given by MACO+ is calculated taking into account the Lexical Model.

Finally, the tagging process is carried out by Dynamic Programming Decoding (DPD), taking as input the output tokens of MACO+ with their respective lexical probabilities.

3 Experimental Work

In order to test our system, we have used the Spanish corpora CPirápides and LEXESP. CPirápides is a corpus consisting of approximately 800 short written sentences (5 Kw). On the other hand, LEXESP is a multidisciplinary corpus, which contains 5.5 Mw of written material, including general news, sports news, scientific articles, etc. LEXESP presents more complex sentences than CPirápides.

The tagset used for tagging both corpora consists of 62 tags defined in the project PAROLE [Martí98]. These tags do not include morphosyntactic information such as gender and number. Therefore we can not use the agreement and inheritance operators to check morphosyntactic agreement. A subset of LEXESP has been used to learn the language model and the lexical model of the tagger. It consists of 75 Kw manually tagged words. CPirápides and a subset of LEXESP (approximately 3 Kw) has been used to test our tagger (and also the full system). The error rates obtained of the tagging process are 0.8% on CPirápides and 3.0% on LEXESP.

```
Level 1
NSV -> (VMI|VMS|VMC|VMM|VAI|VAS|VAC|VAM)CS(VMN|VAN)(VMG|VMP)?
NSV -> PP?PP?(((VMI|VMS|VMC|VMM)(((VMN|VAN)(VMG|VAG|VMP)?)|(VMG |
VAG))?)|((VAI|VAS|VAC|VAM)(VMP|VAP|VMN|VAN)*(VMG | VAG)?))
NSV -> (VMG|VAG)
NSVI -> (VMN|VAN)(((CC|Fc) (VMN|VAN))* CC (VMN | VAN))?
SADJ -> RG? (AQ | VMP) (((CC|Fc) RG? (AQ | VMP))* CC RG? (AQ | VMP))?
NSN -> ((NP (((CC|Fc) NP)* (CC NP))?) | (NC (((CC|Fc) NC)* (CC NC))?))+
Level 2
SN -> TD (PX|PI) (AQ|VMP)?
SN -> DI (PP|PI|PD|P0)
SN -> (DD|DP|DT|DE|DI|D0|TD|TI|(MC* (CC MC)?)|MO)* (SADJ? | Z? | RG) NSN SADJ?
SN -> (PP|PD|PX|PI|PT|P0)|((DD|DP|DT|DE|DI|D0|TD|TI) P0)
SN -> (DD|DP|DT|DE|DI|D0|TD|TI|MO)* (Z| W | MC* (CC MC)?)
SN -> TD SADJ
Level 3
SPR -> SP TD? (PP|PI) D0
SPR -> SP SN
SPR -> SP SADJ
SPR -> SP NSVI
Level 4
SUB -> (SP? CS)|(SP? TD? PR)
Level 5
SADV -> (SP RG) | (RG RG?)
```

Figure 5: Levels and Patterns defined

The syntactic structures identified in the different levels of analysis have been: Noun Phrase (SN), Verbal Heads (NSV), Prepositional Phrase (SPR), Adjective Phrase (SADJ), Infinitive Heads (NSVI), Adverbial Phrase (SADV), Conjunctions and Relative Pronouns (SUB).

In this work we have defined five levels of analysis. The set of non-recursive patterns of each level is showed in Figure 5. These patterns are an improvement of those used in [Molina99] in order to increase the coverage of the parser.

In order to evaluate the performance of the system and to test the influence of the errors of tagging on the syntactic parser, we have conducted two kinds of experiments. In the first one we have considered as input a text without tagging errors (Manually Corrected Tagging, CT) in order to evaluate only the performance of the syntactic parser. In the second one, we have used APOLN taking as input the output of our tagger (Tagger Output, TO). In **Table 1** and **Table 2**, we summarize precision[1] an recall[2] rates per pattern obtained in the experiments defined above for Cpirápides and LEXESP test corpora respectively.

We have revised the errors produced by the parser in both experiments. In CT experiments (Experiment 1 and 3) the most common errors are due to identify incorrectly adverbial locutions (SADV) and some adjective phrases (SADJ). Also, some compound nouns and compound adjective phrases are incorrectly attached. We can not solve this kind of ambiguity with the information provided by the lexical tags, because it would be necessary other information sources (e.g. semantic or contextual information). Obviously, in TO experiments (Experiment 2 and 4) the performance of the parser decreases because of the tagging error. The most usual errors of the tagger are due to the following confusions: adjectives and nouns, adjectives and verbs and adjectives and adverbs. This influences mainly the precision and recall rates of the SADJ pattern. In the Appendix we show some examples of usual errors that were observed.

Corpus CPirápides		**NSV**	**NSVI**	**SN**	**SUB**	**SPR**	**SADJ**	**SADV**
Experiment 1	**Precision (%)**	99.6	100.0	99.0	100.0	99.0	100.0	95.2
CT + APOLN	**Recall (%)**	97.7	100.0	98.5	94.7	98.7	66.7	95.2
Experiment 2	**Precision (%)**	99.6	100.0	98.5	94.1	98.3	66.7	94.7
TO + APOLN	**Recall (%)**	97.1	100.0	97.6	84.2	95.5	66,7	85.7

Table 1: Precision and recall for CPirápides

Corpus LEXESP		**NSV**	**NSVI**	**SN**	**SUB**	**SPR**	**SADJ**	**SADV**
Experiment 3	**Precision (%)**	97.6	100.0	98.9	100.0	96.3	77.6	100.0
CT + APOLN	**Recall (%)**	97.6	100.0	97.5	100.0	95.8	80.9	97.6
Experiment 4	**Precision (%)**	94.3	85.7	92.0	99.4	92.9	64.2	95.0
TO + APOLN	**Recall (%)**	94.7	100.0	89.4	100.0	92.4	72.3	92.7

Table 2: Precision and recall for LEXESP

[1] Precision: total number of correct tags given by the parser / total number of tags given by the parser * 100

[2] Recall: total number of correct tags given by the parser / total number of tags in reference corpus * 100

4 Conclusions and Future Work

In this paper we have presented an incremental partial parser based on finite-state machines, that is able to identify syntactic structures on unrestricted text. The experiments performed have given good results identifying phrases, although the amount of available test set (supervised and syntactically parsed text) could be scarce to provide statistically significant results.

Moreover, we are developing a parsing system that allows us to completely parse an unrestricted corpus. The system use APOLN as first step of processing. The entire parsed corpus could be useful as an information source for treating linguistic phenomena and for developing inductive methods based on corpus.

5 Acknowledgements

This paper has been supported by the Spanish CICYT project TIC97-0671-C02-01/02

References

[Abney96] Abney, S. "Partial Parsing via Finite-State Cascades". In *ESSLLI'96 Robust Parsing Workshop*. 1996.

[Abney97] Abney, S. "Tagging and Partial Parsing". *Corpus-Based Methods in Language and Speech processing*. S. Young y G.Bloothooft Eds. Kluwer Academic Publishers 1997.

[Aït-Mokhtar97] Aït-Mokhtar, S. Chanod, J.P. "Incremental Finite-State Parsing". In *Proc. of the Fifth Conference on Applied Natural Language Processing*. Washington D.C., USA, 1997.

[Carmona98] Carmona J., Cervell S., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J. "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text". In *LREC'98*, 1998.

[Chanod96] Chanod, J.P., Tapanainen, P. "A Robust Finite-State Parser for French". In *ESSLLI'96 Robust Parsing Workshop*. 1996.

[Ejerhed88] Ejerhed, E.I. "Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods". In *Proc. of Second Conference on Applied Natural Language Processing*. ACL, 1988.

[Marcus93] Marcus, M.P., Santorini, B. Marcinkiewicz, M.A.. "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics* n° 19, 1993

[Martí98] Martí M.A.,Rodríguez H.,Serrano J. "Declaración de categorías morfosintácticas". Doc.ITEM n°2.UPC, UB, 1998.

[Molina98] Molina, A. Moreno, L. "Técnicas de Análisis Parcial en Procesamiento del Lenguaje Natural". Internal Report DSIC-II/30/98. Septiembre 1998.

[Molina99] Molina, A., Pla, F., Moreno L., Prieto N., "Incremental partial parser of unrestricted natural language sentences". *In SNRFAI99,* Bilbao 1999.

[Pla98] Pla, F. Prieto, N. "Using Grammatical Inference Methods for Automatic Part-of-Speech Tagging". In *LREC'98,* 1998.

[Prieto92] Prieto N. & Vidal E. "Learning Language Models through the ECGI Method". *Speech Communic.*, 11, 1992.

[Rulot89] Rulot H., Prieto N. & Vidal E. "Learning accurate finite-state structural models of words through the ECGI algorithms". *Proc. of International Conference on Acoustic and Speech Signal Processing*, 1989.

Appendix: Examples of usual parsing and tagging errors.

- Wrong Tag: NC →AQ

[Esta DD envidia NC] SN [de SP la TD que PR] SUB [todos PI] SN [somos VAI] NSV **[víctimas NC] SN y CC [verdugos AQ] SADJ** [se PP ha VAI cebado VMP] NSV [en SP tí PP] SPR [singularmente RG] SADV , Fc ...

... [víctimas NC y CC verdugos NC] SN ...

- Wrong Tag: RG →AQ

[No RG] SADV [estoy VMI] NSV **[seguro RG] SADV** [de SP que PR] SUB [las TD generaciones NC] SN [de SP hoy RG] SADV [sepan VMS] NSV [del SP poderoso AQ influjo NC] SPR ...

... [seguro AQ] SADJ ...

- Wrong Tag: VMS → AQ

Y CC [entre SP sus DP disfraces NC] SPR [no RG] SADV [es VAI] NSV el TD **[menos RG] SADV [frecuente VMS] NSV** el TD [de SP la TD ideología NC] SPR . Fp

... [el TD menos RG frecuente AQ] SN ...

- Wrong Tag: AQ → NC

[Un TI cuarto NC] SN y CC [último AQ punto NC] SN [me PP parece VMI] NSV [interesante AQ] SADJ [destacar VMN] NSVI [a SP este DD respecto NC] SPR (Fap [como CS] SUB [lo PP he VAI hecho VMP] NSV [en SP mi DP mencionado VMP libro NC El_Continente NP vacío AQ] SPR) Fcp . Fp

[Un TI cuarto AQ y CC último AQ punto NC] SN ...

- Syntactic Error: Compound Noun

[Para SP evitarlo VMN] SPR , Fc [conspicuos AQ fascistas NC] SN , Fc [monárquicos AQ] SADJ [a SP la TD violeta NC] SPR , Fc [marxistas AQ] SADJ **[de SP provincias NC y CC católicos NC wojtilianos AQ]** SPR [se PP han VAI venido VMP confabulando VMG] NSV [en_eso RG] SADV [que CS] SUB [ahora RG] SADV [llaman VMI] NSV [pacto NC] SN [a SP la TD griega NC] SPR ...

... [de SP provincias NC] SPR y CC [católicos NC wojtilianos AQ] SN ...

- Syntactic Error: Compound Adjective

[No RG] SADV [estoy VMI] NSV [seguro RG] SADV [de SP que PR] SUB [las TD generaciones NC] SN [de SP hoy RG] SADV [sepan VMS] NSV **[del SP poderoso AQ influjo NC] SPR , Fc [intelectual AQ y CC moral AQ] SADJ** , Fc [que CS] SUB [tu DP magisterio NC universitario AQ] SN [ejerció VMI] NSV [sobre SP la TD disidencia NC juvenil AQ] SPR ...

... [del SP poderoso AQ influjo NC , Fc intelectual AQ y CC moral AQ] SPR ...

- Syntactic Error: Adverbial Locution

Y CC [es VAI] NSV [sobre SP esto PD] SPR [sobre SP lo PP] SPR [que PR] SUB , Fc [resuelto VMI] NSV [el TD breve AQ paréntesis NC] SN [de SP mis DP cartas NC] SPR , Fc [me PP gustaría VMC ahondar VMN] NSV **un TI [poco RG más RG] SADV** . Fp

... [Un_poco_más RG] SADV ...

About Arabic Electronic Dictionaries and their use in Rule-based NLP Methods

CHIRAZ BEN OTHMANE –ADNANE ZRIBI

ABSTRACT

We present in this paper a condensed description of a research that is carried on several years along[a]. The main concern of this research is electronic dictionaries and their use in NLP applications. Arabic is not the only language we worked on, but it is the main one. We propose here to solely consider one aspect: dictionary construction. How to construct an electronic dictionary for Arabic? What type of entries should we try to collect in it? What other information should it include? When built, how to describe its contents? What other results can we expect from a dictionary construction system? What influence do different dictionaries have on NLP applications?

Introduction

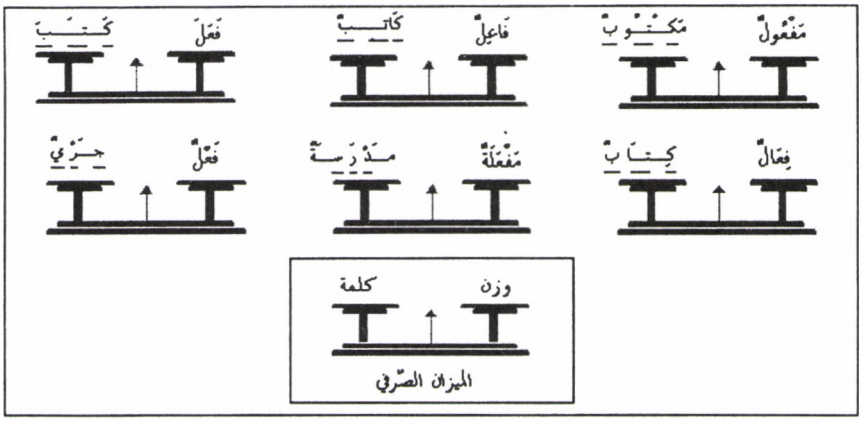
Our goal was the construction of an electronic dictionary for the Arabic language. This dictionary, that have to contain all Arabic inflected forms, is to be used in diverse NLP applications that deal with Arabic. Our methods had to be rather practical i.e. we aimed to really construct the dictionary and not to just study theoretical aspects of such a construction.

Generate Arabic inflected forms automatically ?

Arabic is a very derivational language. Relatively well-defined rules are here to govern word construction. On the first sight, one would be tempted to generate the entire dictionary by automatic methods.

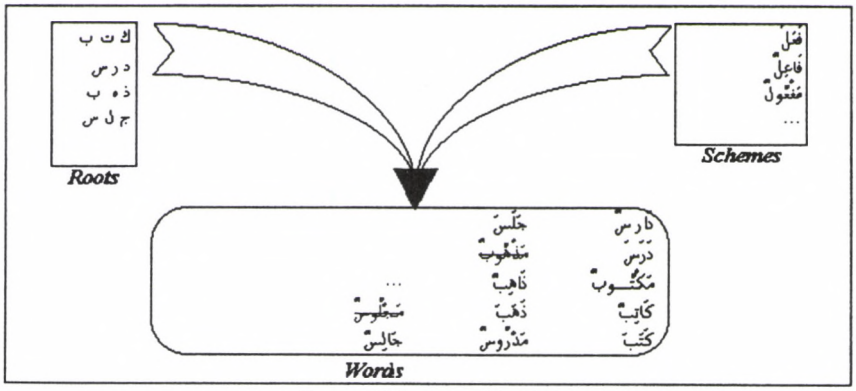
---

[a] This paper is based on research carried on in the French C.N.R.S. and in Université Paris-sud Centre d'Orsay. A related doctorate thesis was accomplished on December 1998 by BEN OTHMANE, C. "*From lexicographic synthesis to misspelled words detection and correction in Arabic*", PH.D. Thesis, University of Paris XI, ORSAY, 1998



Words = Roots + Schemes.

It rapidly appears that building all inflected forms of Arabic by automatic means is simply impossible. Multiple examples show that human intervention in the generation process is inevitable.



Roots + Schemes = Words ?

Of course, this manual intervention permits to deal with exceptions but it also (and more often) permits to select among many behaviors words can have when inflected. This leads us to say that the generation system we set up is semi-automatic. It holds some automatic procedures and other manual ones. Automatic procedures are simply computer programs to be run. The so-called "manual procedures" permit, by using "intelligent" software tools, to introduce; to verify and to fix collected linguistic data. We implemented in these programs a set of rules and algorithms in order to reach three goals:

- Minimize the manual interventions
- Make them as simple and easy as possible
- Validate their results

We incidentally point here the importance of ergonomic computer programs in the research field of natural language processing.

The general method we adopted to build our dictionary can be summarized as follows:

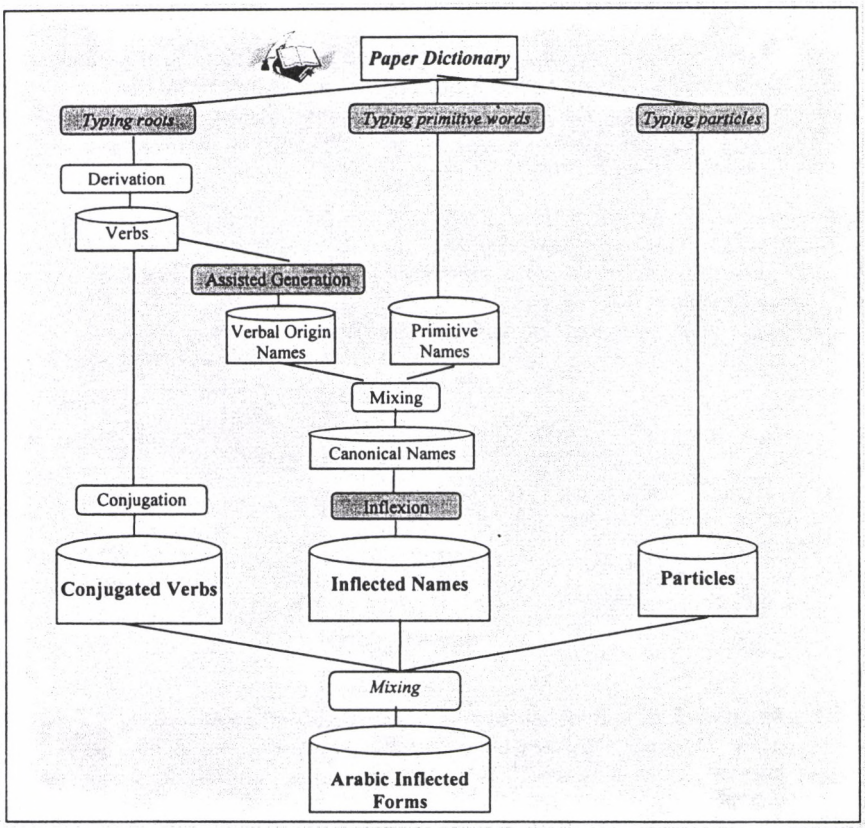
Manually typewrite a minimum quantity of data;
Assign linguistic, statistic and other rules;
Build algorithms inspired by theses rules;

Generate automatically when possible the greatest number of entries and linguistic information.

Implementation

The generation system we build is composed of three parts. One is concerned with verbs, the second treats nouns and the third is there to collect particles. A final step, called mixing step, permits to assemble the results of these three components. At the end of the entire process, we get a dictionary of about 600 000 Arabic inflected and non diacritized forms. For each form, you find a set of linguistic information such as vowels, grammatical class, lemma, etc. The number of diacritized forms represented by the dictionary is more than 1 600 000.

The following schema shows the whole system architecture:



General architecture of the generation system.

In the above schema, procedures represented by gray boxes are semi-automatic. This means that manual intervention is required in these procedures.

Results

The achievements we have reached through the construction of an electronic lexicon for Arabic gave us other results that are as important as the dictionary itself. One result is the importance of the linguistic observations we were be able to perform. Another result is the statistics made possible by the electronic dictionary. In fact, after each of the three components of the generation system, and after the final mixing step, series of statistical studies help us to achieve a quite precise perception of the collected data. Some of these statistics come to corroborate assertions announced by Arabic linguists. Others bring these assertions more precision and even perfect them.

An ultimate result of the system described above, is algorithms, rules and linguistic data that compose the system itself. Definitely, these tools are precious means to comprehend and master Arabic forms generation.

Arabic agglutinated words

We explored a completely distinct experimentation lane. No more satisfied to collect Arabic inflected forms, we try to generate all Arabic agglutinated words. Having such a dictionary would make Arabic in the same situation of other languages such as English or French. These languages use dictionaries containing forms as they appear in texts. But is such a generation possible? How much entries would we have? What would be the size of the dictionary? What consequences will this have on algorithms and applications?

We did construct such a dictionary. It holds about 140 million diacritized entries. If we consider only the words without storing linguistic data with them, we need about 600 Mb of disk space. The hugeness of the needed disk space makes using such a dictionary in text analysis methods quite difficult. However, we give here a comparison of two analysis methods: one using inflected forms and an analysis grammar, the other using only agglutinated words lexicon.

	Inflected words + Grammar	Agglutinated Words
1. Simplest analysis algorithms	☐	☒
2. Homogeneous methods for multiple languages	☐	☒
3. Fastest analysis	☐	☒
4. Smallest space	☒	☐
5. Easiest lexicons update	☒	☐

Conclusion

The problem we faced is the non existence or the non availability of electronic dictionaries for the Arabic language. In this paper, we gave a brief description of a system we set up which goal was to generate all inflected forms in Arabic. This dictionary is now constructed. We are using it in multiple applications that need morphological and syntactic text analysis. Many other results have been attained by this generation. Among them, a whole set of statistics made possible by the availability of Arabic dictionary on electronic form.

Tagging and Conversion of a Bilingual Dictionary for XeLDA, a Xerox Computer Assisted Translation System

TADEUSZ PIOTROWSKI

The paper discusses issues related to adapting a conventional medium-size bilingual English-Polish dictionary to the needs of a prototype computer-assisted translation system, developed by Xerox, within the EC-funded Copernicus project STEEL (Developing Specialized Translation/Foreign Language Understanding Tools for Eastern European Languages). The issues will be: conversion into SGML format, based on TEX tags, insertion of appropriate part of speech labels for English, and description of multiword expressions by means of IDAREX formalism. The dictionary had innovative features and the focus will be on how standard procedures were adapted to those features.

The project and the tasks

The paper will discuss the issues related to one stage in the EC-funded project COPERNICUS STEEL: Developing Specialized Translation/Foreign Language Understanding Tools for Eastern European Languages, still in progress (1997-1999). Both Western and Central European partners participate in the project, from France (Xerox), Poland (Warsaw University and the author), the Czech Republic, and Germany. The most important tool to be developed is a prototype of an intelligent electronic bilingual dictionary for comprehension of specialist texts (English-Polish and English-Czech), based on Xerox's XeLDA format for bilingual SGML dictionaries, in which the results of Xerox's experience in various European projects, notably in the COMPASS project, will be used. As a result of the work of the partners the prototype of an English-Polish dictionary is already functioning. My paper will focus on adaptation of an existing bilingual English-Polish dictionary to the system.

The system has two 'standard' features, and one non-standard (Poirier 1998). As to the two standard features, first, the system requires that the dictionary should be in the SGML format. Second, the lookup system of the dictionary uses part of speech (POS) labels to map parts of speech identified by the tagger in text, and thus to restrict the scope of search in the entry. The third feature is non-standard: the system uses a grammar, IDAREX (Karttunen et al. 1997; Segond and Breidt 1995), for encoding multiword expressions, in particular idioms, which are coded as a regular sequence of words and linguistic tags. The codes are used by the LOCOLEX engine to recognize multiword expressions in text and to return possible translations. These features had to be incorporated in the dictionary that was to be adapated for the project.

The dictionary that was used for the system was a medium-size general English-Polish and Polish-English dictionary, with approximately 44,000 entries. Specialist entries were to be added o the second stage of the project. For the system only the English-Polish part was used (with some 22,000 entries), as at the onset of the project there were no computational tools available for Polish, in particular no grammatical tagger, or tokenizer, etc.. The dictionary was a typical bilingual dictionary in that it had rich microstructure and many space-saving devices. It was written for conventional book publication by me and Zygmunt Saloni in the early 1990's, and was later revised (Piotrowski and Saloni 1997; henceforth *Nowy*), and was, to my knowledge, the first in Poland to use a system of structural tags, based on TEX typographical tags. The tags were quite idiosyncratic, but, in the course of revisions, they were refined and were believed to be fairly unambiguous, so that they were considered suitable to be used for the conversion into the SGML format.

What is important for the project and for this paper is that the dictionary, as compared to other dictionaries, had some features that, against the prevailing tradition, can be called innovative. The dictionary was written with a human user in mind and the authors made an attempt to make it as useful as possible, providing extensive user-friendly information in a restricted space. Therefore the user was to be exclusively a Pole, an intermediate learner of English, who would use the English-Polish part as a comprehension dictionary. From these assumptions a number of features resulted, which made it not easy to adapt the dictionary to the task. Thus, first, part of speech (POS) labels were not used at all, and, second, the dictionary was strictly a translation dictionary. These two features will be discussed in what follows.

In the printed dictionary POS labels were not used, as, first, they are rarely utilized by the user, second, the Polish equivalent most often coincides in its part of speech with the English lexeme, therefore POS information is redundant, and, third, in those cases where there were differences textual examples were usually added to help the users reach the most adequate equivalent. More, however, has to be said about the translation nature of the dictionary (this was based on theoretical views in Piotrowski 1994).

Most bilingual dictionaries do claim to be translation dictionaries, and most often they are useful

for translators. Their structure, however, is most often that of a typical defining dictionary, in that the structure of the entry is built on intralinguistic senses (meaning) of the given lexeme. The semantic structure of one language provides thus a skeleton to which the equivalents are added. For comprehension, and from the point of view of the economy of space, this system has many drawbacks. First of all, there are numerous repetitions of the same equivalents. Second, the entry uses structuring based on semantic distinctions of which the user is not aware: a Pole does not know the semantics of, say, an English verb *make* when he or she reaches for a dictionary, therefore meaning discrimination based on English semantics is useless for him/her. Third, this approach makes an entry long and difficult to access. All this is shown in the abbreviated example below (a detailed discussion of these issues can be found in Piotrowski 1994).

Table 1 A typical bilingual dictionary entry

Collins Pons English-German Dictionary (1986), entry **make** (abbreviated)
make ... 1. (*produce, prepare*) machen ... **2.** (*do, execute*) ... machen ... **3.** (*cause to be or become*) machen ... **5.** (*earn*) ... machen ... **8.** (*equal*) ... machen ...etc.

In contrast, in the *Nowy* dictionary the primary criterion throughout is substitutability in appropriate context of an English lexical segment by a Polish segment: the arrangement of equivalents and their selection are guided by this criterion. Also idioms are defined with this criterion in mind: an idiom is a string of segments that cannot be translated adequately or naturally using any of the equivalents listed in the dictionary for the given word, or which can be translated adequately only in a certain lexical context. Thus, for the *Nowy* dictionary an idiom is not a linguistic concept but rather a lexicographic one: it is relative to the description of words in the dictionary. Therefore, in the *Nowy* dictionary there are idioms that would be treated as regular meanings in other dictionaries.

However, as a result, it is very rarely that an equivalent has to be repeated in the dictionary. In the entry **make**, with its nine lexical senses, no equivalent appears twice. One equivalent is used twice, for six equivalents, in one of the three syntactical senses of the verb (as in *make sb happy*). Further, the dictionary decontextualized the pair L2-L1 to the maximum extent (providing of course relevant wider context where deemed necessary). Examples of this can be found in the section on IDAREX code insertion. Therefore the dictionary, though compact, can be adequately used to process even advanced texts.

POS labels insertion

For XeLDA POS labels are very important, as it starts search for equivalents by identifying the part of speech of a word in text first. As might be expected, it was found that for some classes of words dictionaries of English are very strongly in disagreement, for example for prepositions or adverbs. Therefore the set of POS labels chosen for insertion in the dictionary was minimal and fairly conservative, predominantly based on those found in the *Oxford Advanced Learner's Dictionary of Current English*, which were also quite traditional.

As already said, the *Nowy* dictionary does not provide POS labels, and it was necessary to insert them in an appropriate place in the structure. The placing of POS labels itself was not trivial, as between the equivalents and the name of the entry there can appear various elements, for example a marker, with a counter, of one of the major parts of an entry — entries in general were subdivided on the basis of their formal properties: pronunciation or grammar of the headword, either its inflection or syntax: thus *can* would have two parts, one for *can, could, could*, the other for *can, canned, canned*. Long entries were divided on the basis of their syntactic properties: thus, **make** has

three parts, two with verbal uses, one for regular equivalents, the other for equivalents based on syntactic patterns, and the third part with the nominal uses. In general, however, in short entries, with regular inflection, the dictionary did not make any structural differentiation between various parts of speech.

What was more important, however, was that the translation nature of the dictionary made it difficult to insert POS tags, as in dictionaries of that type equivalence extends across syntactic categories. Thus, though the part of speech most often coincides between an English lemma and its Polish equivalent, there were also numerous troublesome points, for example in adjectives (English adjectives often have to be translated by Polish adverbs), prepositions (whose equivalents can belong to various parts of speech, or even be non-lexical, e.g. one of the equivalents of *of* is a hyphen and a word-order rule), etc. A typical example of the lexicographer inserting the labels was that she chose them on the basis of the syntax of the Polish equivalent, not on the basis of the syntax of the English lexeme.

Conversion into SGML

The original TEX tagging, though primarily to be used in typesetting and printing the dictionary, actually used structural description of the dictionary as tags for change of font. The main reason was that such tags were mnemonically easier to identify for authors than purely typographic tags. Further, they provided the necessary flexibility in searching and altering relevant fields in the dictionary and changing fonts globally, with the desired sensitivity. There were very few tags used purely for font, most often {\pbf} — bold; and these were used for the most heterogeneous and marginal categories of structure, for which it was difficult and perhaps unnecessary to devise separate tags. In general, however, the tags were found to be unambiguous.

Table 2 Some tags in the *Nowy* dictionary

Tag	Description
\dezambig	Disambiguating information in the English-Polish part
\dyz	Disambiguating information in the Polish-English part
\fleksja	Inflectional form of the name of the entry
\haslo	Name of the entry (entry word)
\idiom	English idiom in the English-Polish part
\kateg	Category (indicates abbreviations)
\kwalif	Label; indicates also metalinguistic information, etc.
\lacz	Prepositional collocation
\nr	Counter tag
\pbf	Bold
\pelny	Full form of an abbreviation, acronym or contraction
\pidiom	Polish idiom in the Polish-English part
\podhaslo	Name of the subentry
\psl	Italics
\wariant	Variant of the name of the entry, or cross-reference from a variant to a name of the entry
\wym	Pronunciation

Conversion into SGML format was done by colleagues from Warsaw University, Katarzyna Głowińska and Marcin Woliński, under supervision of Janusz Bień for computing and Tadeusz Piotrowski for lexicography. Actually it was done in two stages: first, description of a grammar of the dictionary and mapping this onto the structure of XeLDA was carried out, and, second, the conversion itself was done. They started the task by using LexParse, software developed at Tubingen to parse dictionaries. In their initial experiments they used entries in the letters *q, m, o, p*, and scored

success in over 80% of the analysed entries of the respective letters. Afterwards, however, they abandoned this tool as LexParse does not have the capacity to modify the structure of the input dictionary, it is also fairly slow, and they developed their own procedures in Prolog. Actually, LexParse is more versatile than the procedures devised for the task, yet the tools designed specifically and precisely for conversion into SGML were considered more useful.

The tagging and structure of the dictionary were robust enough, so that conversion could go on. At present 95% of the entries can be converted automatically. All of them are validated correctly by the XeLDA DTD. A high number of the remaining ones are, unfortunately, high-frequency words, like *have, make,* etc., that have very complex entries. There are also 'cross-reference entries', for instance for inflectionally irregular words, which simply refer the user to regular entries. Such entries are not accounted for in the XeLDA structure, because the system handles inflection on its own, and there is no need for such entries.

Table 3 Conversion of TEX tags into SGML

Entry **awake**	
TEX tagging	SGML tagging
{(\haslo awake) [{\wym {\e}{\`}weik}], {\fleksja awoke} [{\wym {\e}{\`}wouk}], {\fleksja awoken} [{\wym {\e}{\`}wouk{\e}n}] {\pos V} {\nr 1.} budzi"c (si"e) {\pos Adj} {\nr 2.} czuwaj"acy ({\dezambig nie "spi"acy}) $\diamond$ {\nr 3.} {\idiom be awake} <idarex> be V: :awake </idarex> nie spa"c {\nr 4.} {\idiom fully awake} <idarex> :fully :awake </idarex> rozbudzony }	<entry> <headword> <spl>awake</spl> </headword> <hwinfo> <pronunciation> <phonetic>{\e}{\`}weik</phonetic> </pronunciation> </hwinfo> <syntactic> <senseinfo> <pos>V</pos> </senseinfo> <semantic> <subsense> <trans>budzić się</trans> <trans>budzić</trans> </subsense> </semantic> </syntactic> <syntactic> <senseinfo> <pos>Adj</pos> </senseinfo> <semantic> <subsense> <trans>czuwającynie śpiący</trans> </subsense> </semantic> <semantic> <subsense> <idiom>be awake</idiom> <idarex>be V: :awake</idarex> <trans>nie spać</trans> </subsense> </semantic> <semantic> <subsense> <idiom>fully awake</idiom> <idarex>:fully :awake</idarex> <trans>rozbudzony</trans> </subsense> </semantic>

Entry **awake**	
TEX tagging	SGML tagging
	</syntactic> </entry>

IDAREX codes insertion

The IDAREX grammar is what makes the dictionary different from other ones, as it allows XeLDA to correctly identify multiword expressions in text. Those multiword expressions are not only idioms but any multiword lexemes, such as phrasal verbs and compounds. For speeding up the analysis of idioms all idioms in the *Nowy* dictionary were extracted and processed by XeLDA, so that most of them were described by the IDAREX coding. Afterwards these codings were used as basis for interpretation of the idioms in the *Nowy* dictionary.

There were two difficulties in this task. One was that there were strings treated as idioms in the *Nowy* dictionary which, from the point of view of English semantics, are not idioms (such as be *apologetic about/for*, *be applicable*: many adjectives have to be adequately translated by Polish verbs, and then the English adjective has to be used in a verbal frame). They are idioms from the particular standpoint of the *Nowy* dictionary. Second, quite often the form an idiom had in the *Nowy* dictionary was different from that recognized by XeLDA. This made recognition of the idiom impossible or difficult. The *Nowy* dictionary attempts to list all idioms in their canonical form, i.e. in a maximally decontextualized form (with verbs, when inflected, in the infinitive, etc.). Thus, the *Nowy* has an idiom *into thin air*, which is properly *to disappear* or *vanish into thin air*. However, from the point of view of comprehension, the abbreviated form *into thin air* is quite adequate because the dictionary has also entries for *vanish* and *disappear*, and the user can interpret the idiom in a componential way. Another problem was variants of idioms in the dictionary (cf. *as *it is/it turns out/things stand**, which is used instead of 'as it is, as it turns out, as things stand'). In all such instances the form of the idiom was rewritten for IDAREX coding so that it was the same as that recognized by XeLDA. All variants were spelled out as separate idioms. The lexicographers used only very rudimentary computing tools for this task, a text editor with macro instructions, and they had to read entry by entry.

Conclusions

Problems in the execution of tasks in conversion, tagging and annotation of the *Nowy* dictionary in the STEEL project resulted from several sources. These can be grouped under such headings as: tools, ambiguity of the printed dictionary, and dictionary-specific nature of the used solutions.

As for tools, there was a huge difference between computer specialists and lexicographers. It was certainly interesting and perhaps understandable that computer specialists, involved in the conversion, wanted to automate their task as far as possible, therefore they developed an instrument specifically designed to perform the task in hand in preference to an instrument that had a wide range of application (LexParse) which were not, however, immediately useful. The Prolog procedures were a one-off tool, as they can be used for conversion only of the structure of the *Nowy* dictionary. They probably cannot be re-used in other tasks of this nature. Thus, the specialists wanted to avoid manual intervention. Lexicographers, on the other hand, preferred — or, perhaps, were able — to work by hand, using only the most rudimentary tools.

As is common also in work of this type, it turned out that a dictionary meant for human users has many inconsistencies from the point of view of computational requirements on one hand. On the other hand, the dictionary is also redundant. These inconsistencies resulted from various causes. One, which was to be expected, was ambiguity of the TEX-derived tags system, which were meant to be used by a typesetting system. Another was the natural tendency to economize on space in a printed

dictionary, which was also to be user friendly; this resulted in the use of numerous graphic devices, such as slashes to introduce variants, asterisks, to mark the boundaries of variants, etc. All such notations had to be rewritten in full. Finally, the dictionary was redundant because it had information which the system could handle on its own. This related first of all to English inflection, both within the entries and in cross-references, English lexical variants (British, American, etc.), within the entries and in cross-references. XeLDA's DTD simply did not account for that type of information. Any information of this type had to be removed or moved to other fields in the dictionary. In other words, the dictionary had to be often re-written according to the specifications of XeLDA.

Yet the authors of XeLDA sugested (Poirier 1998) that XeLDA's solutions were universal, applicable to other dictionaries, with of other languages. Yet it has to be remembered that these solutions were based on the structure of a specific dictionary (*Oxford-Hachette English and French Dictionary*), and they were considered to be universal enough. The *Oxford-Hachette* was treated as a paradigmatic case for a whole class of bilingual dictionaries, which was natural. From many points of view the *Oxford-Hachette* is an admirable and excellent dictionary, yet in its structure it is also very traditional. It is true that commercial dictionaries are traditional, or, in other words, repetitive and imitative (cf. Sinclair 1984: 5; and the discussion in Piotrowski 1994: 14), and that most of them try to use the same solutions for fear of opposition or criticism from the public. In many points the *Nowy* dictionary, as discussed, is unlike the *Oxford-Hachette*, and the categories derived from the English-French dictionary were ill-adjusted to it. Yet it also is representative of a — smaller — group of bilingual dictionaries (see examples in Piotrowski 1994). The results show that the claim about the universality of the concepts about bilingual lexicography, embedded in XeLDA, are exaggerated. It also, however, stresses the need for a more detailed description of bilingual dictionaries.

References

Karttunen, L. et al. (1997). "Regular Expressions for Language Engineering", *Natural Language Engineering*, 1-24

Oxford Advanced Learner's Dictionary of Current English, V edition. 1995. Oxford: OUP

Oxford-Hachette English and French Dictionary, 1994. Oxford: OUP

Piotrowski, T. (1994). *Problems in Bilingual Lexicography*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego

Piotrowski, T, Z. Saloni. 1997. *Nowy Słownik Angielsko-Polski i Polsko-Angielski*. Warszawa: Wilga

Poirier, H. (1998). *XeLDA Bilingual Dictionary Format*, Xerox Research Centre Europe Document. Grenoble

Segond, F. & Breidt. (1995). "IDAREX: Description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis", *Lexicomatique et Dictionnairiques*, Actes des IVe Journées Scientifiques du reseau ' Lexicologie, Terminologuie, Traduction' de l'UREF, Lyon, Septembre 1995

Sinclair, J. (1984). "Lexicography as an Academic Subject". In *LEXeter '83 Proceedings. Papers from the International Conference on Lexicography at Exeter*. (Ed. R.R.K. Hartmann), pp. 3-12. Tübingen: Max Niemeyer Verlag

INTEX Tutorial Notes

MAX SILBERZTEIN

1. Introduction

INTEX is a linguistic development environment that allows users to build large-coverage Finite State descriptions of Natural Languages and apply them to large texts (several dozen million words in real time).

Several modules of INTEX have been available since 1992 under NextStep; INTEX has been fully integrated in a graphical interface since 1996 (release 3.0), at which point it began to be distributed to research centers as a linguistic development tool. INTEX has been ported to Windows 95-NT in 1997, as INTEX 4.0. Latest version is 4.12.

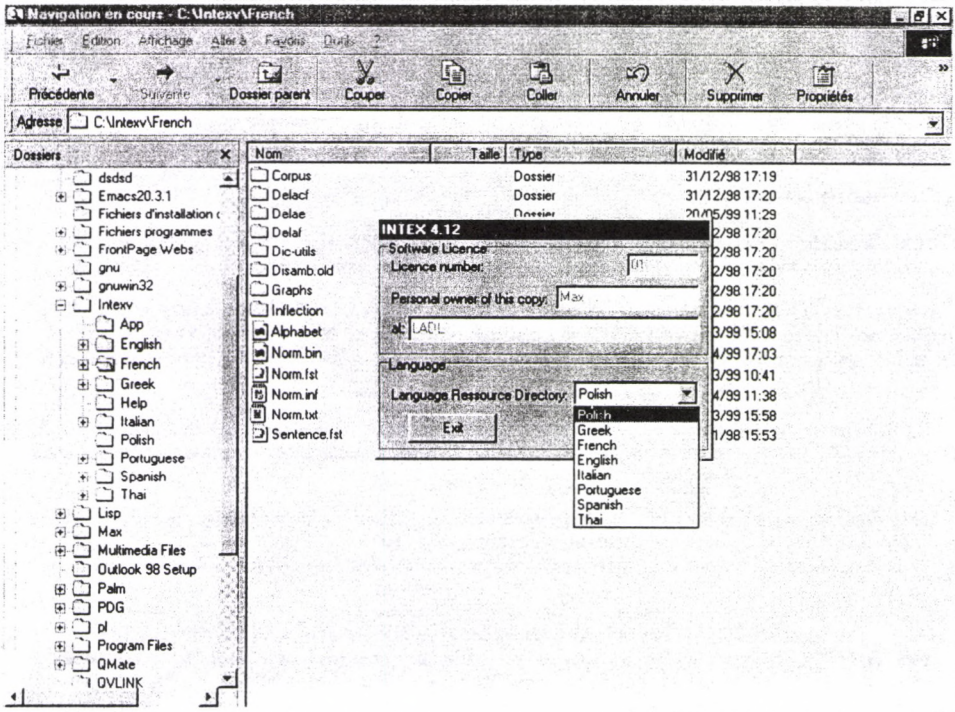
INTEX is based on the work performed at the Laboratoire d'Automatique Documentaire et Linguistique (LADL), founded in 1967 by Prof. Maurice Gross. The goal of the LADL is to build a large-coverage description of Natural Languages by using 3 sets of tools:

- *Electronic dictionaries* and Finite State descriptions of the vocabulary and the morphology of Natural languages;
- *Local grammars* to identify frozen, semi-frozen and phrases in texts;
- Transformational rules described in *Lexicon-Grammars* to extract elementary sentences from complex sentences.

One important aspect of INTEX is that texts, dictionaries and grammars are all represented by Finite State Transducers (FSTs). Therefore, all the operations the user performs via the graphical interface are translated into a small number of elementary operations on FSTs (about 30 programs). For instance, applying a set of dictionaries to a text is performed by constructing a union of the dictionaries' FSTs, then applying the resulting FST to the text FST; removing lexical ambiguities in the text is performed by computing the intersection between a grammar FST and the text FST, etc.

2. Launching INTEX

The first operation consists of selecting the directory where the linguistic data is stored: alphabet of the language, preprocessing dictionaries and Finite State Transducers (FSTs), dictionaries for simple and compound words, FSTs representing the inflectional and derivational morphology of the language, FSTs used to remove lexical ambiguities, utilities for the maintenance of the dictionaries and the grammars.

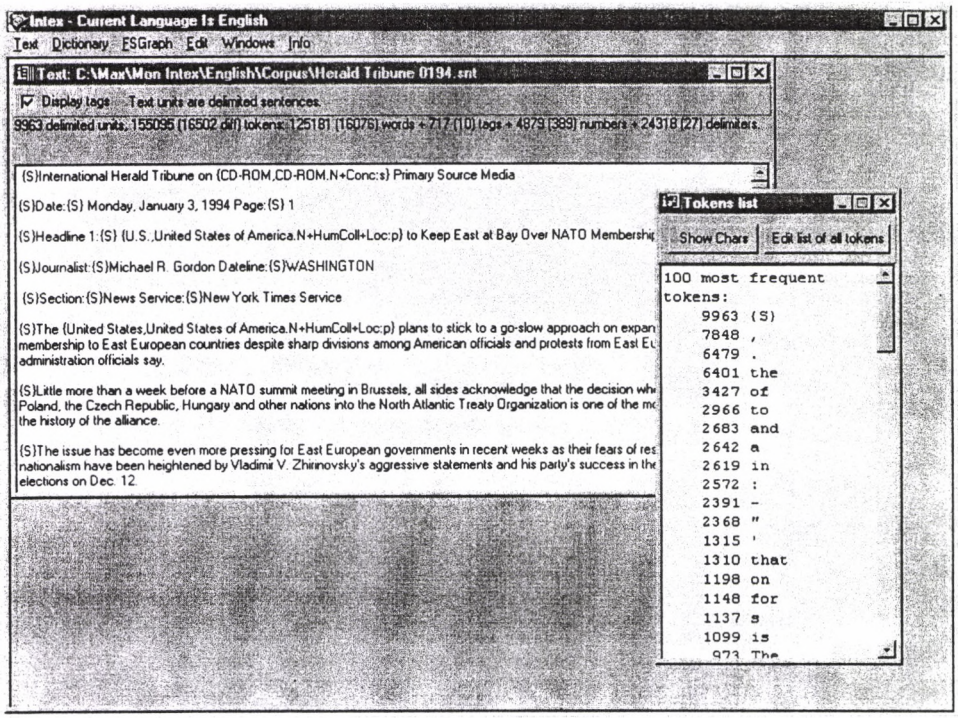


INTEX has no hard-coded information on languages: the alphabet, the dictionaries and the grammars are stored in each language directory.

3. Opening a text (Text->Open)

INTEX processes two types of texts:

- Windows ANSI files are considered as raw texts; they are considered as sequences of units delimited by the NEWLINE character; they do not contain any linguistic data;
- INTEX-formated files are texts that contain some linguistic tags (at least sentence delimiters);



INTEX processes four kind of tokens: **Words** are sequences of letters; **Tags** are sequences of character between curly brackets; **Numbers** are sequences of digits; **Delimiters** are characters not listed in the Alphabet file.

Tags are used to represent linguistic information. For instance, `{S}` stands for sentence delimiter, `{has,have.V:P3s}` represents the form *has* of the verb *to have*, conjugated in the present, third person singular.

4. Finite State Transducers

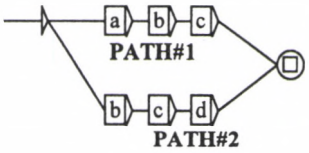
An FST is a device that recognizes some sequences in the input, and associate them with some outputs. Typically, input sequences are sequences of characters or sequences of words in the text; outputs are tags representing some linguistic information. An FST has the form of a graph that starts with an initial state, and ends with a terminal state. Recognized sequences are the ones that can be spelled by a path that goes from the initial state to the terminal state; Outputs of the FST are produced when a sequence has been recognized. Now, let's see how FSTs are applied to texts by the INTEX system:

R1: FSTs are applied from left to right

When the FST has matched one sequence of the text, it is reapplied after the end of the matching sequence. For instance, consider the following text:

```
z a b c d z
```

If we apply the following FST[1] to this text in REPLACE mode (*FST outputs replace matching sequences*):



we produce the resulting text:

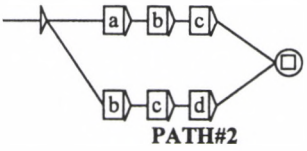
```
z PATH#1 d z
```

The sequence **a b c** matched and has been replaced by the output of the transducer, i.e. **PATH#1**. If we apply the same FST in MERGE mode (*FST outputs are inserted in the text*), we produce the following result:

```
z PATH#1 a b c d z
```

The sequence **a b c** matched, then the output **PATH#1** has been inserted before the character **a**. In these two examples, the sequence **b c d** was not even 'seen' by the system. The sequence **a b c** has priority over the sequence **b c d** because it has matched *before*.

The output of the FST may be the empty string. For instance, the following FST would not modify the text:



[1]. FST inputs are displayed inside boxes, FST outputs are displayed below boxes.

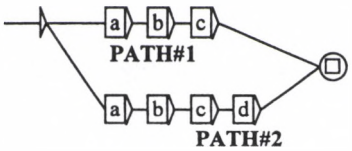
because when **a b c** matched, the empty string has been inserted before **a**; then the FST is being applied at the position at **d z**.

R2: Longest matches have priority over shorter ones

For instance, consider the following text:

```
z a b c d z
```

If we apply the following FST to it:



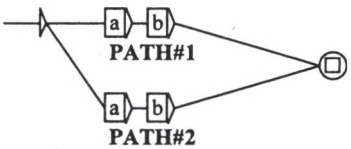
we get the following results:

(*in REPLACE mode*)	`z PATH#2 z`
(*in MERGE mode*)	`z a b PATH#2 c d z`

In other words, the matching sequence **a b c d** has priority over the sequence **a b c** because it is longer. When applying Finite State Automata[2] to texts, users *can* choose to index only shortest matching sequences, only the longest matching sequences, or all matching sequences.

R3: INTEX cannot apply ambiguous FSTs

If one sequence in the text is associated with 2 or more different outputs, INTEX performs an undefined action. For instance, if the following FST is applied to the text **z a b z** in REPLACE mode:



we get one of the two results:

```
z PATH#1 z
```

or

```
z PATH#2 z
```

INTEX *can* apply ambiguous FSTs to texts that are represented by FSTs: each result will be represented as a parallel path in the Text FST. Therefore, ambiguous FSTs *can* be used to disambiguate texts.

---

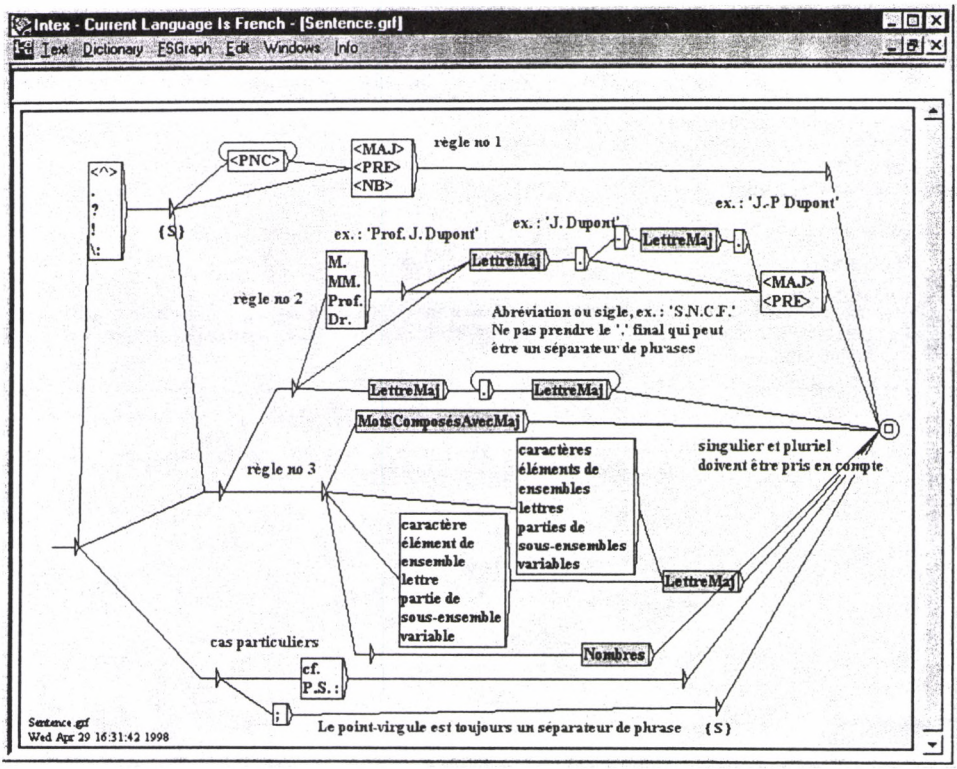
[2] . In INTEX, Finite State Automata are FSTs that produce the empty string.

5. Preprocessing the text (Text->Preprocessing)

Let us go back to the parsing of an raw ASCII file. After having loaded the file, users can preprocess the text, i.e. prepare the text for the linguistic analyses. The preprocessing consists of three operations: identification of sentences, of unambiguous compounds, and of special tokens.

5.1 Identifying sentences

The standard FST `Sentence.fst` (stored in the current language directory) is applied to the text in MERGE mode. Generally, this FST is used to insert the sentence delimiter {S} between consecutive sentences. Further INTEX processing will take this mark into account to process and index every linguistic unit.



Gray nodes refer to embedded graphs; for instance, LettreMaj is the name of a graph that identifies the 26 capital letters A...Z; MotsComposésAvecMaj is the name of a graph that lists all the French compound words that end with a capital letter (e.g. *Vitamine C*).

The graph `Sentence.fst` must be read in the following manner:

- if a period is followed by a word in capital letters, INTEX inserts the sentence delimiter between the period and the word (see on the top of the FST);

- if an single uppercase letter is followed by a period, followed by a word in uppercase (e.g. *J. Dupont*), INTEX does not insert any sentence delimiter;
- compound words that end with an uppercase letter (e.g. *Vitamine C*) may occur at the end of a sentence; the uppercase letter followed by a period must not be mistaken for an abbreviated firstname.

Thanks to the *Left to Right* priority (rule R1), the last processing gets priority over the second processing, which has priority over the first processing. We then get the correct result:

```
J. Dupont comes. {S} Paul eats some vitamine C. {S} Luc also.
```

(*C. Luc* is not processed as *J. Dupont*). Although this FST is not perfect (some systematic errors are due to the use of some English abbreviations in French texts), it process usual French novels and journalistic texts with a high rate of success (>99.5%).

5.2 Identifying unambiguous compounds

The second step of the preprocessing will consist of identifying and tagging unambiguous compound words in the text, this operation corresponds to a look-up of the dictionary Norm.dic (stored in the current language directory).

Ambiguity

Ambiguous words are words that correspond to more than one lexical entry in the dictionaries and the FSTs of the system. Ambiguous compound words are sequences that correspond either to more than one lexical compound entry, such as:

pied noir (*Blackfoot*, Frenchman born in Algeria)
pied noir (*Blackfoot*, American Indian)

or to more than one sequence of lexical (simple or compound) entries, e.g.:

red tape : {red tape,.N} + {red,.A} {tape,.N}

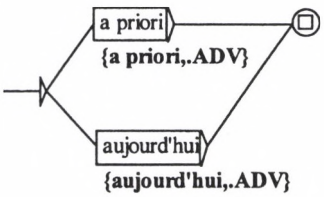
Unambiguous compound words correspond to only one lexical compound entry, e.g.:

a priori : {a priori,.ADV}

Tagging a text consists of replacing its forms by the corresponding lexical entry, written between curly brackets.

One can only tag unambiguous, or disambiguated forms

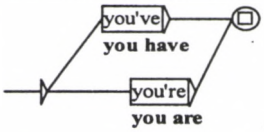
It is desirable to identify and tag unambiguous compound words as soon as possible, in order not to treat their constituents (e.g. '*a*' and '*priori*') as simple words. This operation can be performed during the preprocessing analysis by means of the special dictionary **Norm.dic** stored in the current language directory, or by FSTs applied in REPLACE mode. For instance, the following FST:



would tag the two French unambiguous adverbs. The French dictionary **Norm.dic** lists over one thousand entries.

5.3 Identifying special tokens

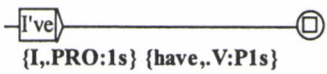
The last stage of the preprocessing consists of identifying and tagging special tokens, such as elided and contracted words, unambiguous abbreviations, etc. This step is performed by applying the FST Norm.fst (stored in the current language directory) in REPLACE mode. For instance, by applying the following FST to English texts:



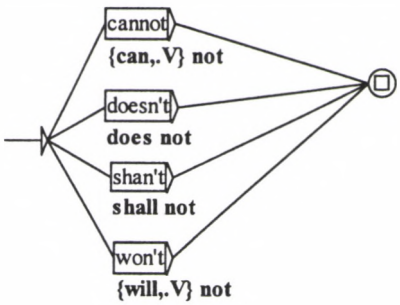
one replaces all the utterances of the sequences *you've* and *you're* by the more explicit sequences *you have, you are*. Such replacements are performed before indexing the text and before consulting the dictionaries. One could perform a similar substitution to replace the sequence *I've* by the form *I have*, but that would lead to artificially add an ambiguity in the text: the sequence *I've* is not ambiguous (it is the pronoun *I*, followed by the verb *have*), while *I have* is ambiguous, as shown in the sentence:

Nelson and Napoleon I have met in ...

(*I* is used as a roman numeral). Thus, it is better to use the following FST:



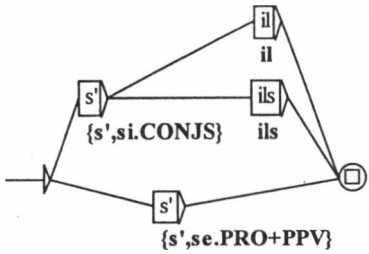
I is the pronoun, first person singular; *have* is the verb conjugated in the present, first person singular. It is also possible to replace one word by a more complex sequence, such as in the following FST:



(the forms *can* and *will* are ambiguous in general; it is better to tag them here).

5.4 Unambiguous sequences of simple words

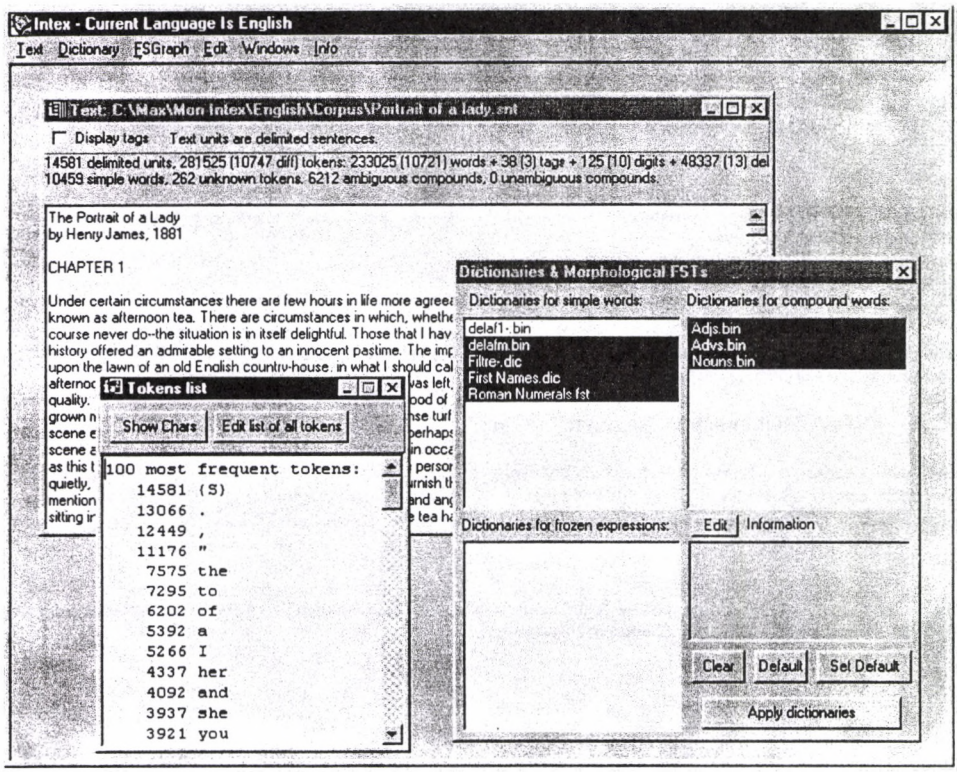
During this early stage of preprocessing, it is also possible to disambiguate a number of grammatical words by replacing them with a tag when they occur in certain contexts. For instance, the following FST:



disambiguates the form s' (Conjunction or Pronoun), according to its immediate right context. Thus, the ambiguous form (*s'* corresponds either to the conjunction *si*, or the pronoun *se*) has been disambiguated.

6. Apply dictionaries and lexical FSTs (Text->Apply dictionaries)

After the completion of the preprocessing stage, the user selects the dictionaries and FSTs used to identify simple words (left column) and compound words (right column) in the text.



The three types of files are:

dic an ASCII file that corresponds to a DELAF-type dictionary;
fst an FST graph
bin a dictionary compacted into a FST represented in a binary file.

6.1 Priority levels

Dictionaries and FSTs are associated with a 3-level priority system used to hide or impose some lexical data:

if a (simple or a compound) form in the text matches a high priority dictionary or FST, the consultation stops;
if not, all 'regular' dictionaries and FSTs are consulted;
if no lexical entry corresponds to the form, low level priority dictionaries and FSTs are consulted.

Giving a high priority to a dictionary or a FST for simple words is generally useful in order to remove 'unpleasant' ambiguities from the general-purpose dictionaries: if one knows that a certain use of a word never occurs in the current text, one can insert the word in a high priority dictionary, without the irrelevant analysis. For instance, the word *la* in French is three-time ambiguous:

{la,le.DET:fs} + {la,le.PRO+PPV} + {la,.N:ms}

la can be either a determiner, a pronoun or the masculine noun (musical note). These three analyses are represented in the general-purpose **Delaf.bin** dictionary. By entering the only first two entries in a dictionary that has priority over the Delaf, one gets rid of the noun (which is not frequent).

Giving a high priority to a dictionary or a FST for compounds is useful to get rid of systematic structural ambiguities when compound words can be shortened. For instance, the following regular expression represents two synonymous variants of the same adverb:

Dans l'intimité (la plus stricte + <E>)

(*in the intimacy*). If the longest variant occurs in a text, INTEX must not parse it as ambiguous, as seen in the following two tagged results:

either the adverb: **{dans l'intimité la plus stricte,ADV}**
or the sequence: **{dans l'intimité,ADV} la plus stricte**

The longest variant has to have priority over the shortest one. Giving priority to dictionaries and FSTs for compounds gives priority to the longest matches.

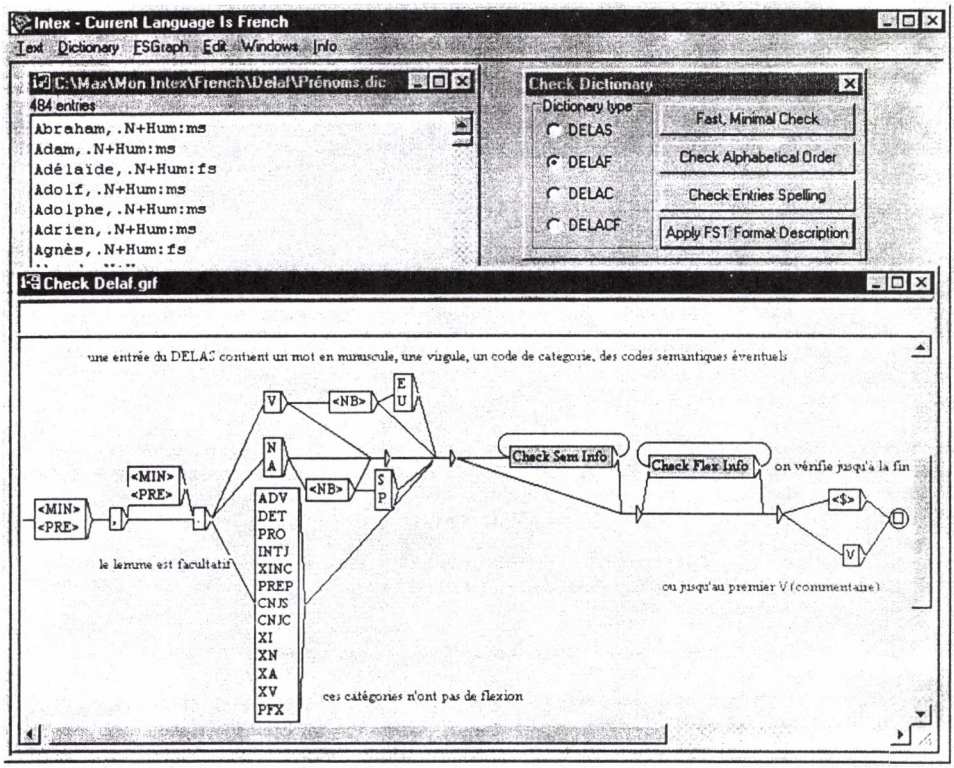
6.2 INTEX Dictionaries (Dictionary->Open)

Dictionaries applied by INTEX to identify simple forms in texts must be in the format of a DELAF; dictionaries used to identify compounds must be in the format of a DELACF. These dictionaries associate text utterances with one lemma and some linguistic information. Here is for instance one entry of the French DELAF:

amuserions,amuser.V+t+4:C1p

the form *amuserions* is associated with the lemma *amuser* which is a transitive (**+t**) verb (**V**) of the syntactic class #4 (**+4**); the form is conjugated in the Conditional, first person plural (**C1p**).

The format of user-defined dictionaries must be checked via the command: Check format[3]:



CheckSemInfo describes the syntactic and semantic features associated with the entry (e.g. **+t** for transitive); **CheckFlexInfo** describes the inflectional codes (e.g.**:ms** for masculine singular).

6.3 From a DELAS to a DELAF (Dictionary->Inflect)

Some dictionaries can be entered by the users directly in the form of a DELAF; typically, dictionaries of first names, abbreviations, etc. However, in most languages, the dictionary of all the inflected forms would be too big to be entered 'manually'. It is then better to manually build a DELAS-type dictionary, and automatically generate the corresponding DELAF.

In a DELAS dictionary, lexical entries are lemmas; each entry is associated with one FST that represents all the corresponding inflected forms. We give below are the first entries of the French DELAS.

V3, A31, N1, N32 and **PREP** are names of inflectional FST files. All the words in a language that have the same set of suffixes, associated with the same inflectional information are associated with the same inflectional FST. For instance, in French, all the verbs that conjugate like *amuser* (*aider*, *voler*, etc.) are associated with the FST **V3**; all the nouns that take an 'e' in the feminine and an 's' in the plural are associated with the FST **N32**, etc.;

[3]. Users may edit the FSTs that describe INTEX dictionaries' format in order to add their own linguistic information.

if a word has no inflection, it must be associated with an FST that has no input and produces no output. For instance, in French, the FST **PREP** is the following (prepositions do not inflect):

Multiple entries in the DELAS

If one word is associated with more than one inflectional class, it must be represented by more than one DELAS entry. For instance:

cousin,N32+Hum
cousin,N1+Anim

The first entry corresponds to a person (= cousin); the noun gets the feminine form *cousine*; the second entry corresponds to an animal (= mosquito); this noun is masculine only. In the same manner, the French DELAS includes the two following entries:

voile,N1+Conc
voile,N21+Conc

the first entry corresponds to a masculine noun (= veil); the second entry corresponds to a feminine noun (= sail). Although both entries are associated with the same distributional class (**+Conc** stands for concrete nouns), and to the same inflection (they both take an 's' in the plural) they must be associated with two different FSTs: one generates the code **m** (masculine) whereas the other generates the code **f** (feminine).

Sometimes, different sets of syntactic or semantic properties may lead to different inflections. For instance, the verb *voler* in the meaning of 'to steal' accepts a passive construction; therefore, the four past participle forms *volé, volée, volés, volées* must be represented, as we see in:

Ces fleurs ont été volées (= these flowers have been stolen)

The verb *voler* in the meaning of 'to fly' is intransitive; therefore, only the invariable form *volé* must be represented. This leads to at least two entries in the DELAS:

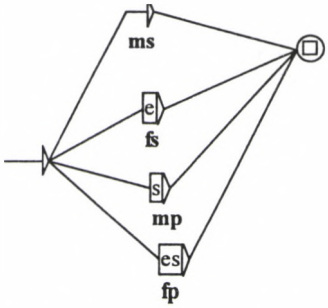
voler,V3U+i
voler,V3+t

V3U represents only one invariable participle form whereas **V3** represents the four forms.

a,N2
a,XI+PMCO
ŕ,PREP
ab,XI+PMCO
abaissé,A32
abaisser,V3+t+11
abaisser,V3+t+32RA
abaisser,V3+t+38L
abaisser,V3E+pi
abandon,N1
abandonné,A32
abandonné,N32+Hum
abandonner,V3+t+32H
abandonner,V3+t+36DT
abandonner,V3+t+38L1
abandonner,V3+t+38LR
abandonner,V3+t+6
abandonner,V3+t+9
abandonner,V3E+pi+7
abandonner,V3U+i+31H
abandonner,V3U+i+35R
abasourdi,A32
abasourdissant,A32
abatage,N1
abats,N2P+Conc
abattage,N1
abattement,N1
abattoir,N1+Conc
abattre,V68+t+32H
abattre,V68+t+32R3
abattre,V68+t+37E
abattre,V68+t+38LD
abattre,V68+t+4
abattre,V68E+pi+35R
abattu,A32
abattu,N1
abbaye,N21+Conc
abc,N2
abcčs,N2+Conc
abdication,N21
abdiquer,V3+t+32R3
abdomen,N1+Conc
abdominal,A76
abdominaux,N2P
abeille,N21+Anl
aberrant,A32
aberration,N21
abętir,V18+t+11
abętir,V18+t+4
abętissant,A32
abętissement,N1
abhorrer,V3+t+12
abîme,N1+Conc
abîmé,A32
abîmer,V3+t+32C
abîmer,V3+t+38LD
abîmer,V3+t+4
abject,A32
abjection,N21
ablation,N21
ablution,N21
abnégation,N21
aboiement,N1
abois,N2P
abolir,V18+t+10
abolir,V18+t+32R2
abolition,N21
abolitionniste,A31
abolitionniste,N31+Hum
abominable,A31
abominablement,ADV
abomination,N21
abominer,V3+t+12
abondamment,ADV+Dadv
abondamment,ADV+PADV
abondance,N1
abondance,N21
abondant,A32
abonder,V3U+i+34L0
abonné,A32
abonné,N32+Hum
abonnement,N1
abonner,V3+t+11
abord,N1
abord,XI+PMCO+Préd
abordable,A31
abordage,N1
aborder,V3+t+32H
aborder,V3+t+32R2
aborder,V3+t+38L1
aborder,V3U+i+35L
abords,N2P
abouti,A32
aboutir,V18U+i+14
aboutir,V18U+i+31R
aboutir,V18U+i+35L
aboutissant,A32
aboutissement,N1
aboyer,V13+t+32R3
aboyer,V13+t+9
abracadabrant,A32
abrégé,N1+Abst
abrégé,N1+Conc
abrégement,N1
abrčgement,N1
abréger,V10+t+32RA
abreuver,V3+t+37M1
abréviation,N21
abri,N1+Conc
abricot,A80+Conc
abricot,N1+Conc
abricotier,N1+Conc
abrité,A32
abriter,V3+t+10
abriter,V3+t+38LD
abriter,V3+t+38R
abriter,V3E+pi+35R
abrogation,N21
abrupt,A32
abrupt,N1+Conc
abruptement,ADV
abrupto,XI+PMCO
abruti,A32
abruti,N32+Hum
abrutir,V18+t+11
abrutir,V18+t+37M1
abrutissant,A32
abrutissement,N1
abscons,A61
absence,N21
absent,A32
absent,N32+Hum
absentéisme,N1
absentéiste,A31
absentéiste,N31+Hum
absenter,V3E+pi+2
absenter,V3E+pi+31H
absinthe,N21+Conc
absolu,A32
...

6.4 Inflectional FSTs

The following FST **N32** associates each suffix of the DELAS entry *cousin* to some inflectional codes:



If nothing is concatenated to the DELAS entry, one gets *cousin*; this form is associated with the inflectional codes ms (masculine singular); if 'e' is concatenated to the DELAS entry, one gets *cousine*; this form is associated with the inflectional codes fs (feminine singular); if 's' is concatenated to the DELAS entry, one gets *cousins*; this form is associated with the inflectional codes mp (masculine plural); if 'es' is concatenated to the DELAS entry, one gets *cousines*; this form is associated with the inflectional codes fp (feminine plural).

A simple process of concatenating the FST **N32** to the lexical entry, and then exploring the FST to generate all the paths will produce the resulting DELAF entries:

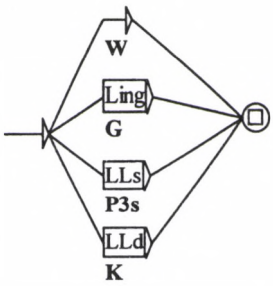
cousin,cousin.N32+Hum:ms
cousine,cousin.N32+Hum:fs
cousins,cousin.N32+Hum:mp
cousines,cousin.N32+Hum:fp

'Delete' operator

Lemmas are not always proper prefixes of their inflected forms. For instance, the verb 'have' gets the following forms:

have, having, has, had

In order to produce the last three forms, one needs to be able to 'delete' characters from the lemma. This operation is performed by adding the 'delete' character **L** (*Go Left*) to the alphabet of the DELAS dictionary. The FST used to produce the four inflected forms is then:



From the DELAS entry: **have,V1**, the process of constructing a DELAF dictionary becomes the following: (1) first we get the four following lines by exploring the FST V1:

have,have.V1:W
haveLing,have.V1:G
haveLLs,have.V1:P3s
haveLLd,have.V1:K

(2) then we spell each inflected form, interpreting the operator **L** as a 'delete last character' command. The final result is then:

have,have.V1:W
having,have.V1:G
has,have.V1:P3s
had,have.V1:K

Stack operators

In theory, the two operators *insertion*, *deletion* are sufficient to represent all kinds of inflection, even highly irregular ones:

avoir LLLLLont => ont

Basically, one can always connect any form to any lemma, by deleting all the letters, and then spelling the form.

The problem I discuss now is a practical one. Consider for instance the following German nouns:

Anrand, Balg, Blatt, Dach, Daus, Fach, Gehalt, Gemach, Geschmack, Gewand, Haus, Inland, Kaff, Kalb, Kraut, Lamm, Land, Mahd, Mann, Mark, Maul, Pfand, Rand, Salband, Wald, Wams...

These nouns have a similar inflected form:

Anränder, Bälger, Blätter, Dächer, Däuser, Fächer, Gehälter, Gemächer, Geschmäcker, Gewänder, Häuser, Inländer, Käffer, Kälber, Kräuter, Lämmer, Länder, Mähder, Männer, Märker, Mäuler, Pfänder, Ränder, Salbänder, Wälder, Wämser...

Basically, the inflection of all these forms is the same: add an ¨ to the *a* two letters before the end, and then add the suffix *er*.

Unfortunately, with the only two operators of insertion and deletion, each of these forms would have to be associated with a different FST:

Anrand LLLänder => Anränder
Balg LLLälger => Bälger
Blatt LLLätter => Blätter

Over 40 almost identical inflectional FSTs would have to be added to the system!

I added two operators to the inflection process: **C** (*Copy*) and **R** (*Go Right*); all the previous inflections are then performed with the same command:

LLLäRCCer
go left three times, insert an 'ä', go right one time, copy the two letters, insert 'er'

The string produced by the FST is processed by four stack operators that take a constant time:

c	insert character 'c' at the end of the form
L	delete last character; push it onto the stack
R	pop the stack
C	copy the character at the top of the stack to the end of the form; pop the stack

Here is how the string **BalgLLLälger** is processed:

Operator		*Resulting form*	*Stack*
B	Insert B	B^	-
a	Insert a	Ba^	-
l	Insert l	Bal^	-
g	Insert g	Balg^	-
L	Push g	Bal^	g
L	Push l	Ba^	l g
L	Push a	B^	a l g
ä	Insert ä	Bä^	a l g
R	Pop	Bä^	l g
C	Pop & Insert	Bäl	g
C	Pop & Insert	Bälg	-
e	Insert e	Bälge	-
r	Insert r	Bälger	-

The result is **Bälger**. In the same manner, the following French verbs are associated with a unique FST V6

acheter, amener, beder, écarteler, élever, peser, semer

The conjugated forms: *achčte, amčne, bčde, écartčle, élčve, pčse* and *sčme* are produced by the command: **LLLLčCC**.

200 FSTs are required for the description of English,
250 FSTs for Italian,
300 for French and Spanish,
450 for Bulgarian,
670 for German.

The construction of DELAF dictionaries takes a time proportional to their length (linear time).

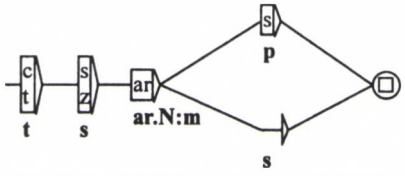
6.5 Resulting DELAF dictionary

a,.N:ms:mp
a,.XI+PMCO
f,.PREP
ab,.XI+PMCO
abaissé,.A:ms
abaissée,abaissé.A:fs
abaissés,abaissé.A:mp
abaissées,abaissé.A:fp
abaisser,.V+t+11:W
abaissant,abaisser.V+t+11:G
abaissé,abaisser.V+t+11:Kms
abaissée,abaisser.V+t+11:Kfs
abaissés,abaisser.V+t+11:Kmp
abaissées,abaisser.V+t+11:Kfp
abaisse,abaisser.V+t+11:P1s:P3s:S1s:S3s:Y2s
abaisses,abaisser.V+t+11:P2s:S2s
abaissons,abaisser.V+t+11:P1p:Y1p
abaissez,abaisser.V+t+11:P2p:Y2p
abaissent,abaisser.V+t+11:P3p:S3p
abaissais,abaisser.V+t+11:I1s:I2s
abaissait,abaisser.V+t+11:I3s
abaissions,abaisser.V+t+11:I1p:S1p
abaissiez,abaisser.V+t+11:I2p:S2p
abaissaient,abaisser.V+t+11:I3p
abaissai,abaisser.V+t+11:J1s
abaissas,abaisser.V+t+11:J2s
abaissa,abaisser.V+t+11:J3s
abaissâmes,abaisser.V+t+11:J1p
abaissâtes,abaisser.V+t+11:J2p
abaissčrent,abaisser.V+t+11:J3p
abaisserai,abaisser.V+t+11:F1s
abaisseras,abaisser.V+t+11:F2s
abaissera,abaisser.V+t+11:F3s
abaisserons,abaisser.V+t+11:F1p
abaisserez,abaisser.V+t+11:F2p
abaisseront,abaisser.V+t+11:F3p
abaissasse,abaisser.V+t+11:T1s
abaissasses,abaisser.V+t+11:T2s
abaissât,abaisser.V+t+11:T3s
abaissassions,abaisser.V+t+11:T1p
abaissassiez,abaisser.V+t+11:T2p
abaissassent,abaisser.V+t+11:T3p
...

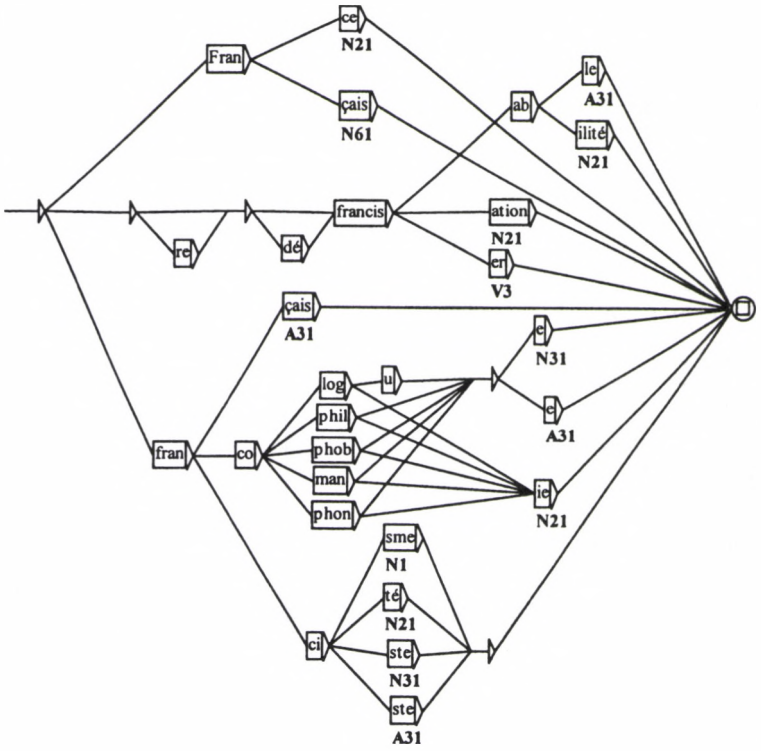
DELAF dictionaries are stored in minimal deterministic Finite State Automata.

6.6 Lexical FSTs

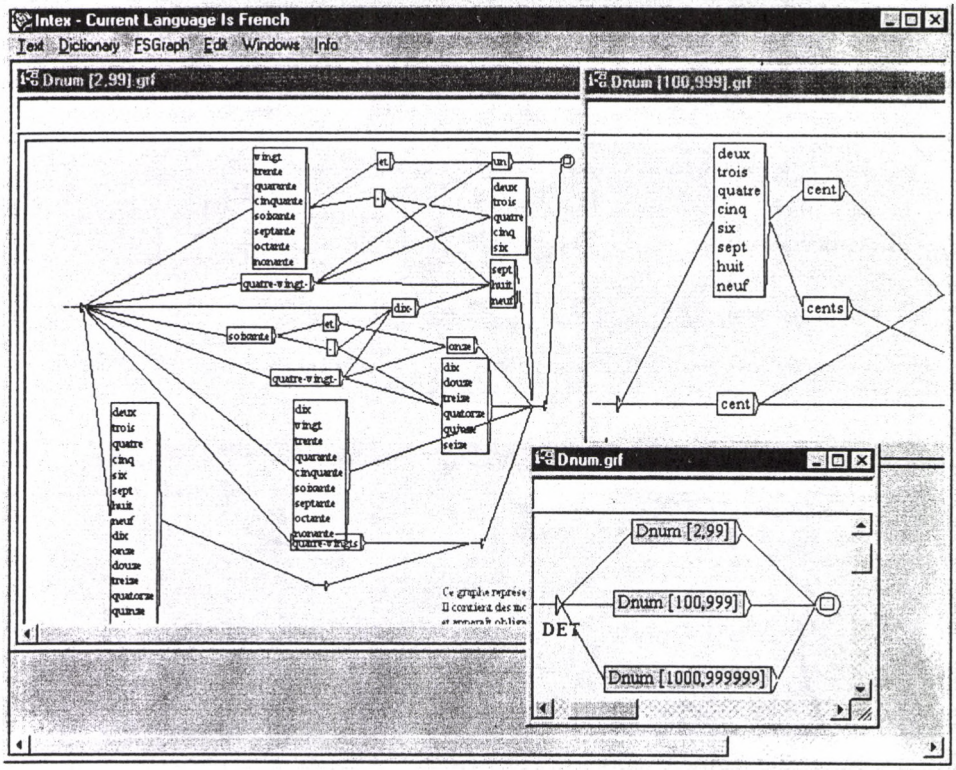
Generally, FSTs must be used to represent infinite sets of lexical entries (e.g. numerical determiners); they are used as well to put together members of a 'natural' family, such as derived forms of a lemma, orthographic variants, synonymous expressions, etc. For instance, here is a FST used to tag the orthographic variants of the noun *tsar* in French:



Here is a FST that represents all the derived forms of the noun *France*:



The following FST Dnum identifies and tag all the numerical determiners from 2 to 999999 written in French:



Since it is possible to embed FSTs in others, one can construct a large library of FSTs that will be re-used in other bigger projects. More than 3,000 FSTs have already been built for various descriptions of French:

technical expressions (e.g. stock market, weather reports, medical statements, etc.),
semi-frozen expressions (e.g. expressions of feelings),
complements of dates, duration, addresses,
etc.

6.7 Text dictionaries

The result of the application of the system dictionaries and FSTs is then displayed.

the list of all the simple forms associated with their lemma and some linguistic data
the list of all the simple forms that have not been found in the selected dictionaries
the list of all compounds, such as 'red tape' (either a person, or a tape)
the list of all frozen expressions, such as 'take ... into account'

If a form is ambiguous, it appears on more than one line. For instance, *abandonné* is ambiguous: either the adjective masculine singular (A:ms), or the noun (N:ms), or the participle of the verb *abandonner* (V:Kms). The sequence *ŕ ce sujet* (= *speaking of this*) can be an adverb; it is considered as an 'ambiguous' compound because the sequence is not obligatorily the adverb, as for instance in:

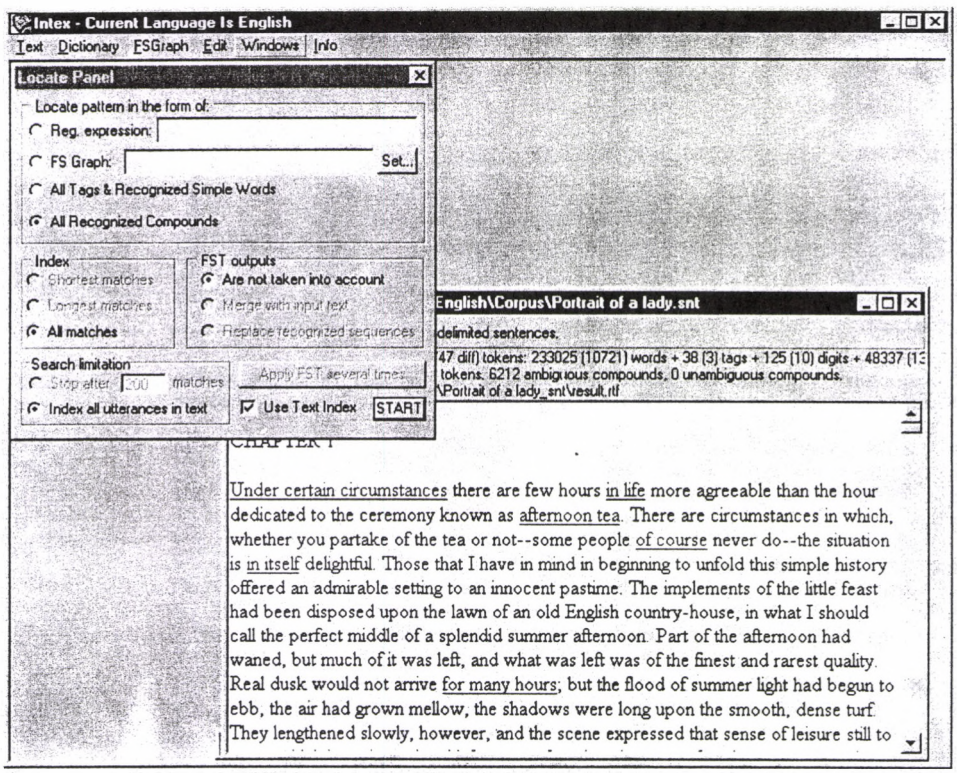
Je pense ŕ ce sujet bien défini
(= I think of this well defined topic)

These four files can be edited, so that the dictionaries used by further processings are closely adapted to the text (one can easily get rid of ambiguities that do not occur in the specific text).

7. Text indexation

7.1 Highlight Compounds in the text (Text->Locate)

All the compounds that have been identified by the consultation of the selected dictionaries are indexed; it is then possible to highlight them in the text (via the *Display indexed sequences* panel).



Generally, the indexation of compound nouns produces a very precious corpus to be used by information retrieval systems, because compound nouns are almost never ambiguous (compared to simple nouns).

7.2 Locate a regular expression in the text (Text->Locate)

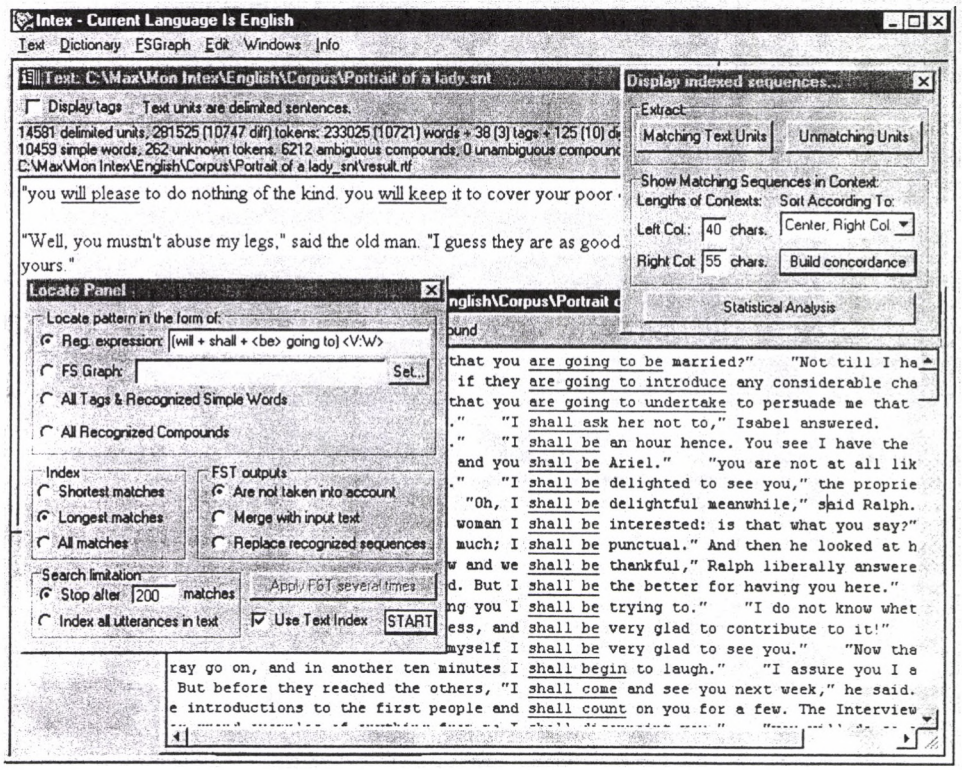
One can index all the sequences of the text that match a particular regular expression. For instance:

<be> going to + (will + shall) <V:W>

<be> stands for any inflected form associated with the lemma *be* (that is, any conjugated form of the verb *to be*); **<V:W>** stands for any form associated with the category V and the inflectional code W (that is, any verb in the infinitive). Any code present in any of the dictionaries is instanteneously useable, for instance:

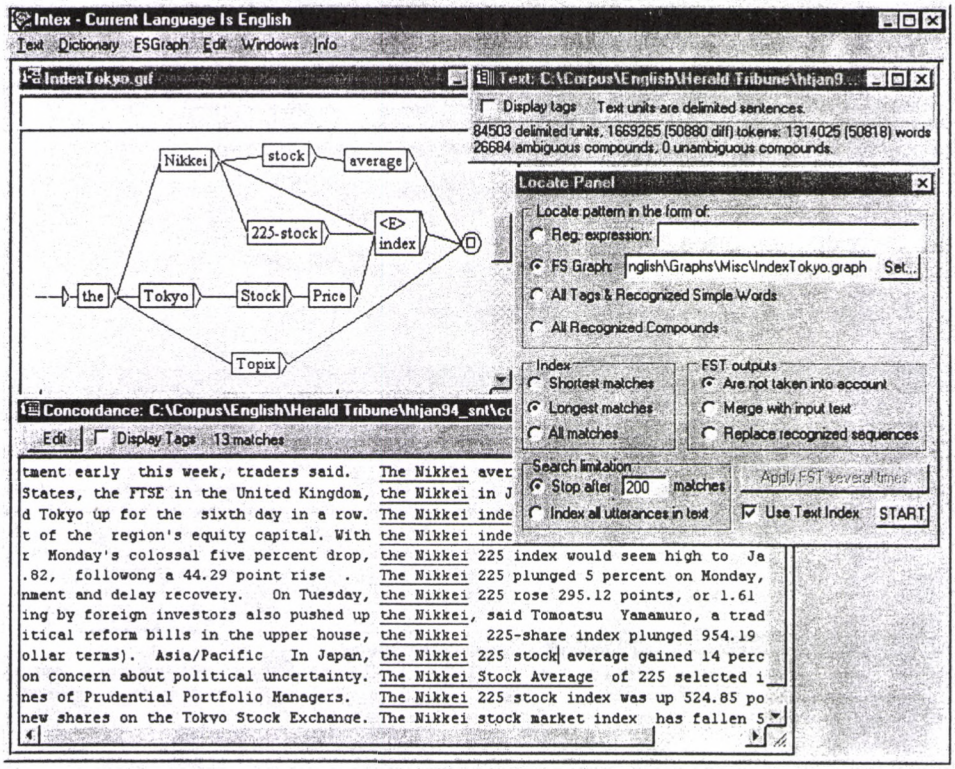
<N+Hum>	Noun, associated with the code Hum (Human)
<N-Hum>	Noun, not Human
<N:fp>	feminine plural Noun
<!PREP>	any word that is not a Preposition

The resulting index can be displayed in the form of a concordance:



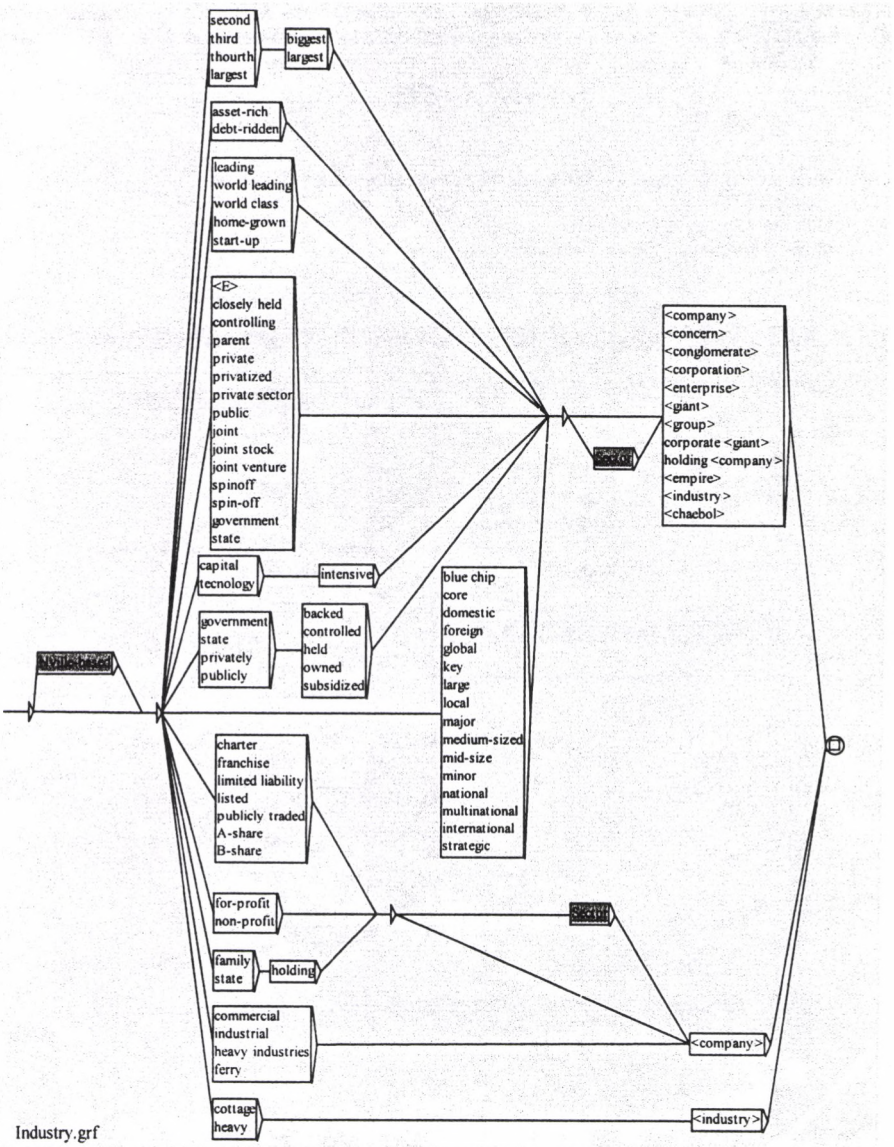
7.3 Index a FST in the text (Text->Locate)

One can index all the sequences of the text that match a particular FST.



Users can choose to index only the shortest matches, the longuest matches, or all matches. FST outputs can be ignored, be merged into the text, or replace matching sequences. With these two options, the FST can be reapplied several times to the text, for instance until no modification has been performed.

One of the grammars used to locate semi-frozen noun phrases that represent companies; when applied to newspapers, the resulting index can be used by information retrieval systems:



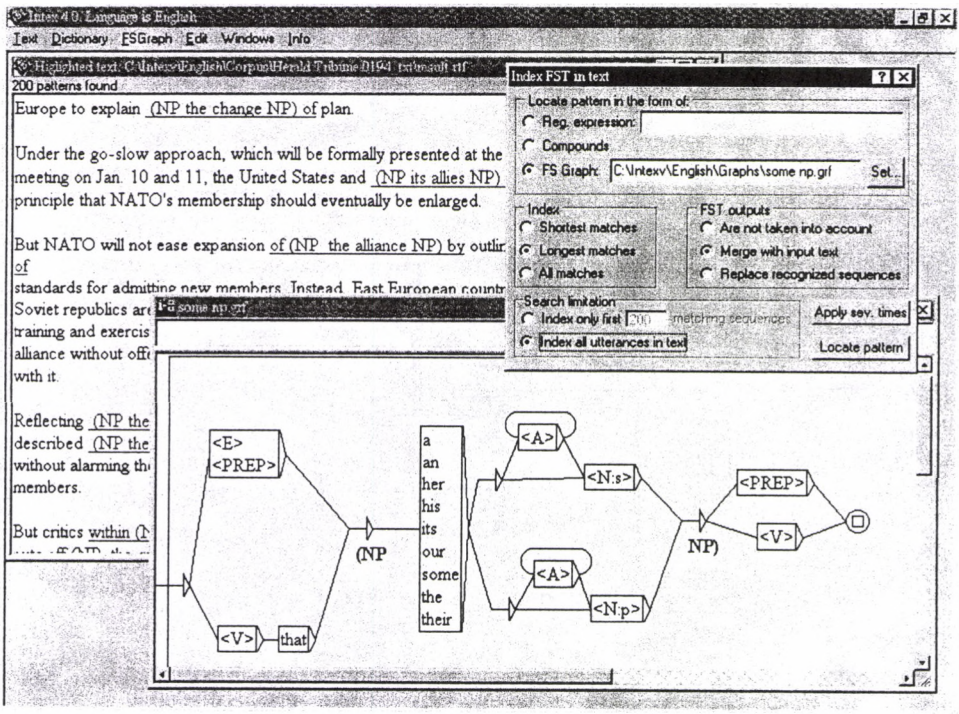
7.4 Various Text Transformations

FSTs can also be applied to texts in order to modify the text itself. For instance, the following FST, if used in REPLACE mode, could be used to remove adverbs from the text (useful for information retrieval systems):

In the same manner, one could design FSTs that would delete expressions like:

I think that, In the same manner, as far as I am concerned,
officials said that, he argued that, it seems doubtfull that, etc.

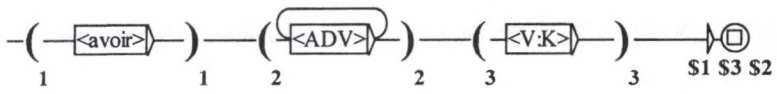
The following FST can be used in MERGE mode to insert parentheses around some noun phrases:



Enhanced FSTs

In INTEX FSTs, outputs and inputs are *synchronized*, that is, the input sequences that match are indexed in the same time as the corresponding output. That feature is essential to the application in MERGE mode (see for instance how parentheses are inserted by the previous FST), or by disambiguation FSTs, where lexical constraints have to be applied to the correct matching forms.

INTEX users are allowed to purposely modify the conditions of this synchronization, by using internal variables to store parts of the matching input sequence. This feature is the same as in the UNIX tool **sed**. Here is one example. The following FST:



is used to extract the adverbs that may be inserted between the auxiliary verb *avoir* and the following past participle. For instance, when the FST is applied to the following text:

Luc m'avait souvent amusé

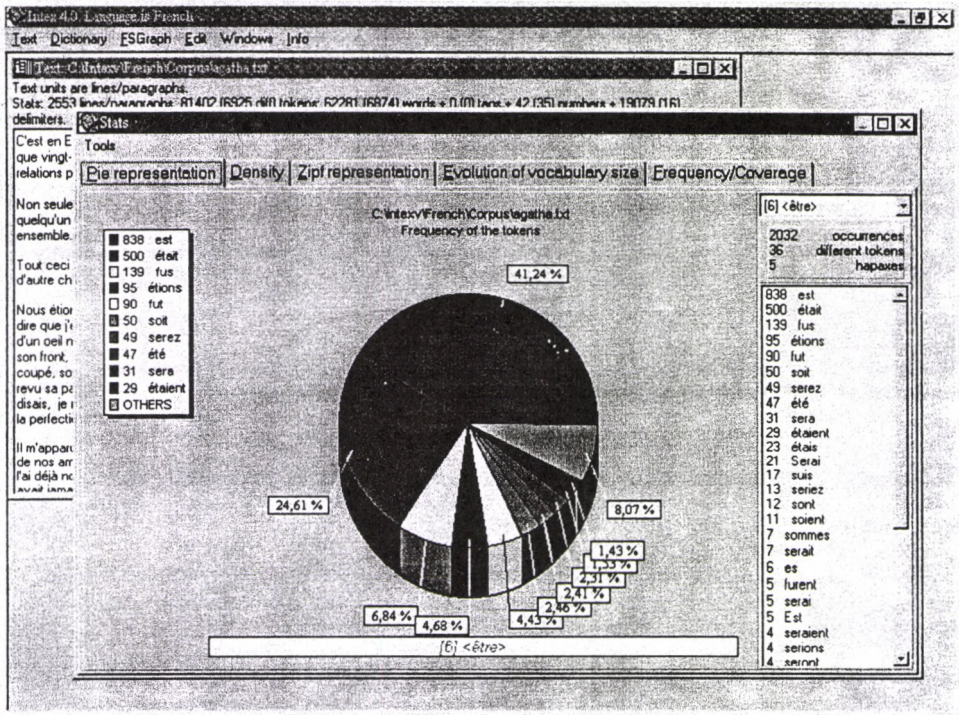
the sequence *avait* is stored in memory slot $1; the sequence souvent is stored in memory slot $2, and the sequence amusé is stored in memory slot $3. Performing the replacement produces the following result:

Luc m'avait amusé souvent

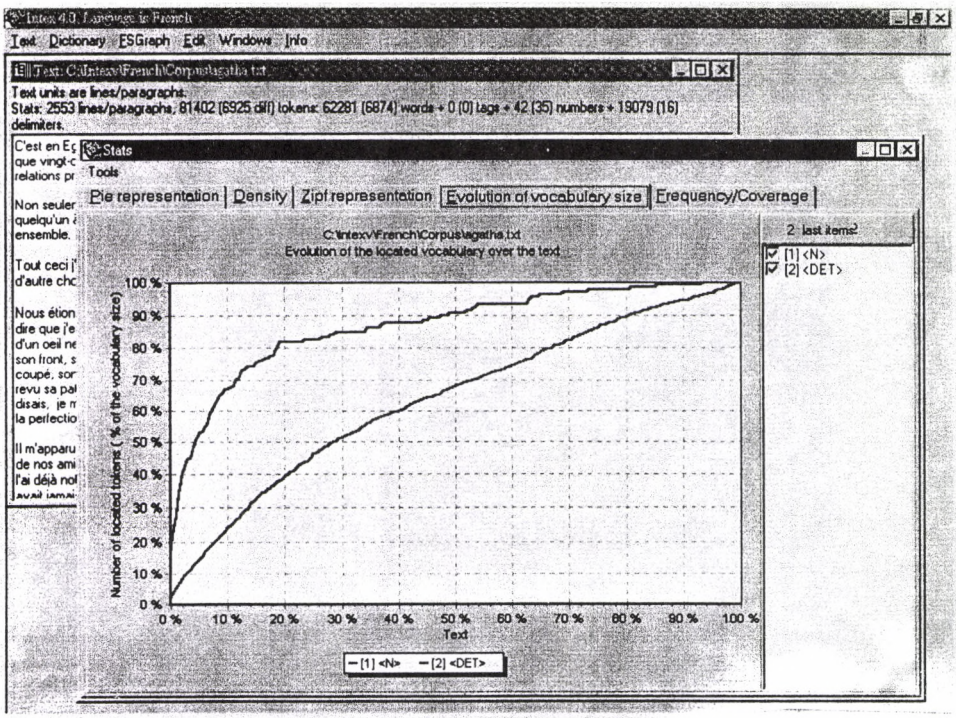
Look at the FST that identifies some French negations in the Norm directory (on the form *ne <V> Neg*).

7.5 Statistical Analyses

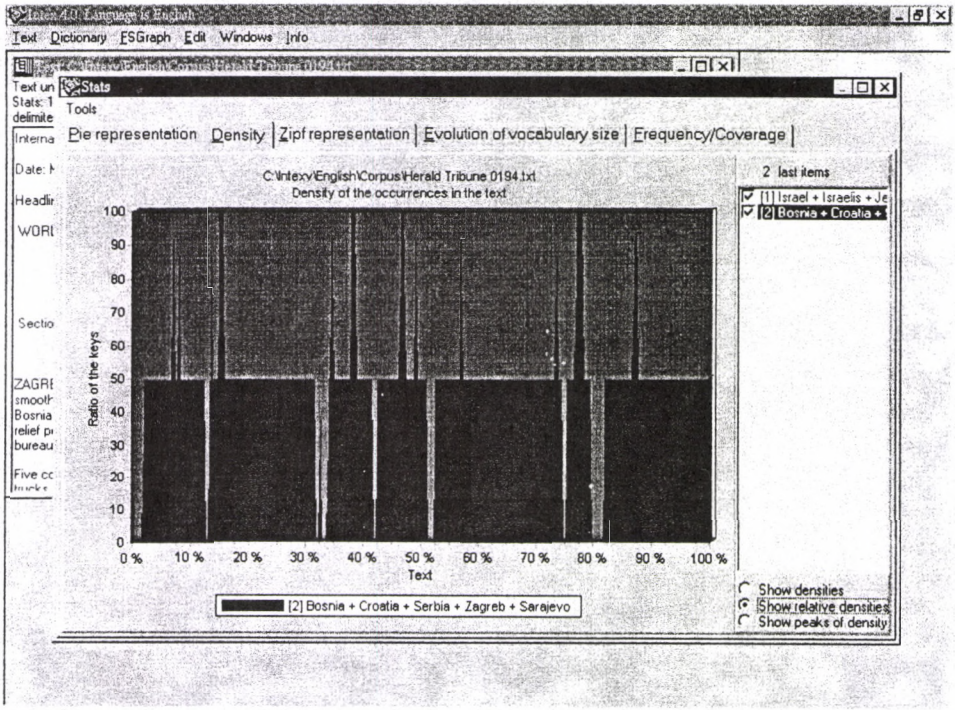
Any index (the index of the compounds, of the sequences that match a regular expression, or a FST) can be studied with a number of statistical measurements. Here for instance, the frequencies of the forms matching the regular expression **<ętre>** (all the conjugated forms of the verb *ętre*) are represented in a pie:



Below, we study the evolution of the number of Nouns (**<N>**) and Determiners (**<DET>**). The proportion of determiners grows much faster than the one of the nouns (which is almost linear).



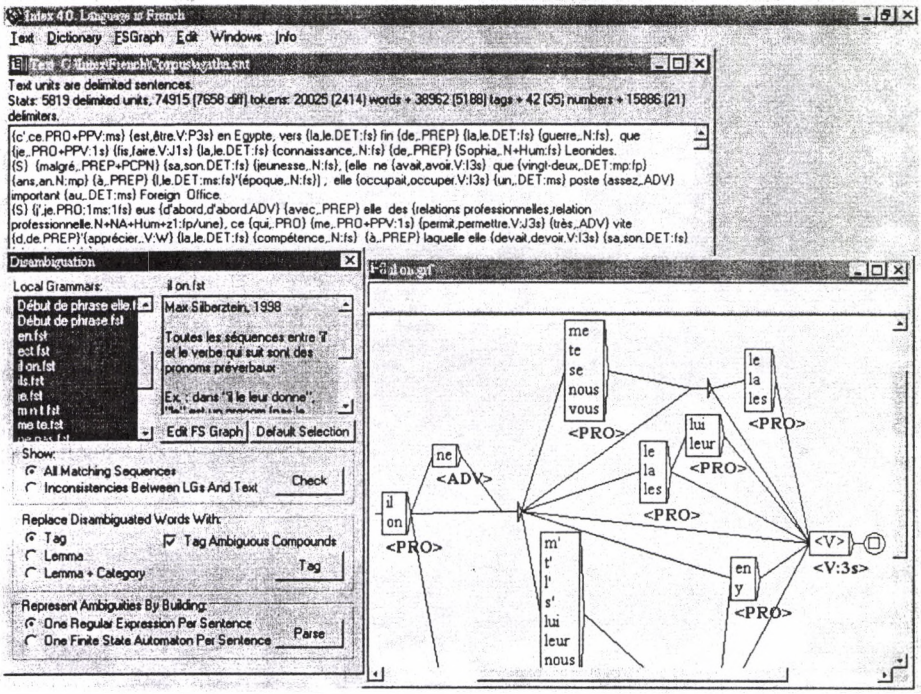
Below we study the density of two regular expressions in the *Herald Tribune*, january 1994:



Peaks correspond to an important local density of the forms that match the indexed regular expression. Various other statistical measures are available.

8. Disambiguation with Local Grammars

Local grammars are FSTs that recognize text sequences (e.g. *il le la donne*), then applies corresponding lexical constraints (e.g. **<PRO> <PRO> <PRO> <V:3s>**) to remove lexical hypotheses.



The result of the application of one or more disambiguating FSTs can be displayed in several ways. INTEX can:

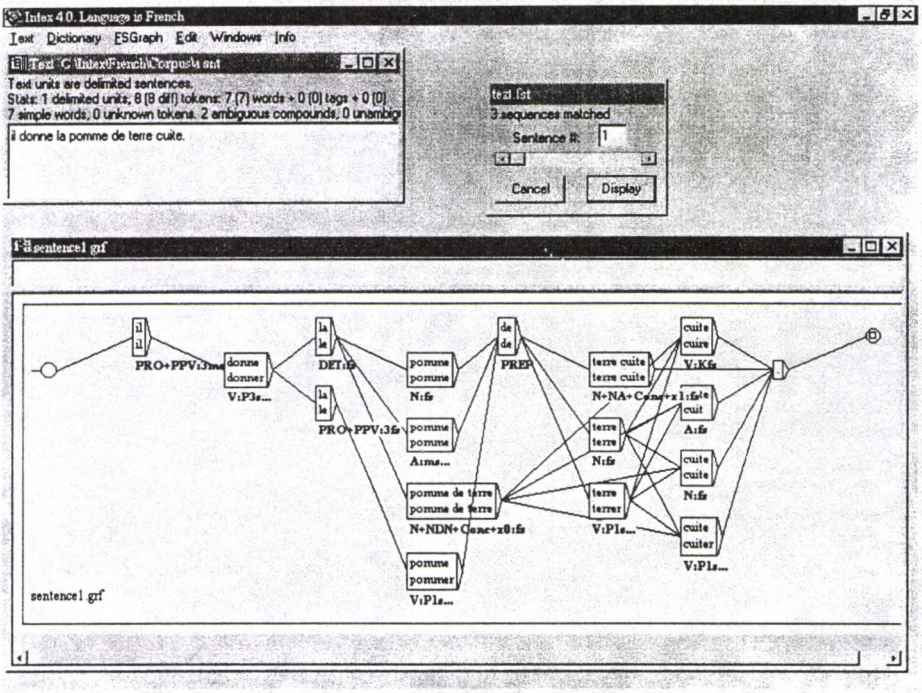
index all matching sequences (i.e. to study the coverage of the local grammars)
index all inconsistencies between selected local grammars and the text (to locate errors in the grammars, or agreement errors in the text)
tag the text, i.e. replace all disambiguated forms by the corresponding lexical entry;
lemmatize the text, i.e. replace all disambiguated forms by the corresponding lemma;
build a regular expression that represents all the ambiguities of the text;
build and display the text in the form of a FST.

9. Lexical Analysis

Internally, each text sentence is represented by a FST. INTEX can display it. For instance, the text:

il donne la pomme de terre cuite

is represented below with all the ambiguities that remain after having applied selected local grammars. Notice how the two ambiguous compound nouns *pomme de terre* (= *potato*) and *terre cuite* (= *clay*) are represented in the FST.



10. Conclusion

10.1 Research Tool

Over 30 research centers are presently using INTEX as a research tool in various domains: computational linguistics, corpus-based linguistics, information retrieval, terminology, literature studies, teaching of a second-language.

Thanks to INTEX, about 60 researchers from all over Europe have started to work within a shared framework, using the same tools and the same methodology to construct large coverage descriptions of a dozen natural languages. This situation is unique in the field of Linguistics.

Large-coverage INTEX descriptions have already been built for Bulgarian (contact Prof. Elena Paskaleva, University of Sofia), English, French and Spanish (Contact Prof. Maurice Gross, Université Paris 7), German (Prof. Franz Günthner, University Maximilian, Münich), Greek (Prof. Ana Symeonides, University of Salonique), Italian (Prof. Annibale Elia, University of Salerne), Polish (Prof. Zygmunt Vetulani), Portuguese (Prof. Elisabete Ranchodd, University of Lisbon), Russian (Alex Kolesov, Academy of Sciences, Moscow). INTEX Dictionaries are being constructed for Korean (Jeesun Nam, University of Seoul) and Old French (Prof. Hava Bat-Zeev, University of Tel Aviv).

10.2 Teaching Tool for Linguistics, Corpus Linguistics and Computational Linguistics

INTEX is a great tool to teach Corpus Linguistics and Computational Linguistics.

Describing the linguistic data included in INTEX (Dictionaries for Simple Words, Description of the Morphology, Dictionaries for Compound words, Dictionaries for frozen expressions, Local grammars for disambiguating texts,

etc.) as well as the technology used to handle this data (Finite State Automata and Transducers) corresponds to a full year course (50 Hours) for Masters students. There are a plethora of projects that remain to be proposed to students, as the description of Natural Languages is not nearly complete!

10.3 Teaching Tool for Second Languages

INTEX is also used to teach French as a second-language; it allows teachers to ask students to locate morpho-syntactic patterns in wide texts (such as 5 years of the newspaper *Le Monde*), to build local grammars for semi-frozen expressions (e.g. how to express a date in French) and to 'play' (edit, correct, apply and test) with some linguistic descriptions.

10.4 New functionalities

INTEX 4.13 will be able to process frozen expressions described in lexicon-grammar tables;
INTEX 4.13 will include a Context Free syntactic parser that will take as its input the text graph, and produce the corresponding derivation trees.
INTEX 4.13 will include a graph debugger.

Starting with *Trauer*. Approaches to Multilingual Lexical Semantics

WOLFGANG TEUBERT

Abstract Multilingual lexical semantics is the core issue of human and machine translation. We discuss the problem of meaning, claiming that so far neither artificial intelligence nor machine translation nor cognitive linguistics has found a viable way to deal with it. We maintain that meaning involves intentionality, i.e. human understanding, and therefore cannot be processed by algorithmic computation alone. Focussing on the semantic field of *Trauer*, *sorrow* mainly in German and English, we show the shortcomings of bi- and monolingual dictionaries if used for translating into a non-native language. We then demonstrate that the same deficiencies are also found in conceptual ontologies. These ontologies do not facilitate the translation of unrestricted general language; rather they add another layer of complication. Therefore we propose corpus linguistics and parallel corpora as a new approach to multilingual lexical semantics, while emphasising the problems still waiting for a satisfactory solution.

1. The problem of meaning

Is it because linguists are generally more timid than philosophers, cognitive scientists or the artificial intelligence community that they tend to leave the problem of meaning to them? One would think they have every reason to attend to it by themselves, for meaning is contained in language and nowhere else (a claim not uncontested even by linguists). If there were no language, there would be no meaning. Only when linguists really cannot avoid the topic do they, or at least some of them, bring in their poor relatives, the lexicographers, making us believe that word meanings are essentially the same as the definitions of word senses we find in dictionaries.

Other linguists delegate semantics to their cousins in artificial intelligence or, on a less mundane level, to cognitivism, or even to the philosophy of mind. One of the reasons why linguists would rather not deal properly with meaning themselves is that they have learned that meaning is, besides form, one of the two essential properties of symbols, and no one has taught them how to deal with symbols. This is why artificial intelligence, cognitivism and a major portion of the philosophy of

1. Introduction

This article is a presentation of the project Lexicó Verbal Bàsic Anglès--Català (LVBAC),[1] a prototypical bilingual English-Catalan lexicon intended as a useful research tool for aspects related to computational linguistics. On completion of the initial stages of the project, the lexicon will have 420 verbs from the basic Catalan vocabulary (*Vocabulari bàsic infantil i d'adults*).

In this article we will establish the methodological basis of our research (see Section 2, *Methodology*), which will be developed at length with a contrastive study (see Section 3, *Contrastivity*) and analysed from a syntactic and semantic point of view (see Subsections 3.1 and 3.2, respectively). In the syntactic aspects, we provide sentences whose translation can be inferred from its syntactic structure divided into syntactic aspects. In the semantic aspects, we give sentences whose translation can be inferred from selectional restrictions applied to the arguments.

Parallel to this project, the lexicon has been implemented and integrated into a program of machine translation from Catalan into English with the purpose of checking how efficient the lexicon may be for machine translation. Section 4 includes a brief summary of this program of analysis.

2. Methodology

The theoretical framework of this study is based on the theories formulated by Gazdar and Mellish (1989). We have made special note of chapter 7, *Features and the lexicon*, where these authors focus on the representation of lexical knowledge and implementation of Prolog lexica. One feature of their model is that they develop syntactic and morphological aspects, particularly the latter, and disregard semantic aspects of verbs.

This study discards the morphological aspects of verbs since these would require individual in-depth analysis for both languages involved[2]. Accordingly, the Project focuses on semantic aspects and we have selected Lenders (1989) and Klavans (1994) as reference theories. In this case, we have introduced the trait *itm* to refer to a lexical item or rather the label we use to define the whole tree. Other traits we have borrowed from the semantic networks include *isa* and *pof*, which refer to a class or type (the boy is animated, the book is inanimated, etc.) and to the relationship of ownership (the arm is part of the body, the column is part of a bulding, etc.), respectively. Although these traits are not applied to the verb but as selectional restrictions of their arguments, they allow us to identify the constructions where these traits can appear. Occasionally, we have used traits which may be have more to do with semantics such as *clas* (countable, uncountable), *fig* (round, flat, rough), etc., but only when we needed them for contrast.

From each verb included in *Vocabulari bàsic* of Catalan, we have constructed a series of examples which are relevant enough to be included on the different syntactic and semantic structures. In order to perform this task systematically, we have departed dictionary definitions and, in some cases, from examples included in dictionaries, particularly *Diccionari de la llengua catalana* (1932) and the monolingual Catalan dictionary by the Institut d'Estudis Catalans. We have also looked up other lexicographic reference works such as the *Gran Enciclopèdia Catalana* and the *Gran Larousse Català*.

---

[1] The LVBAC project was started in 1996 and is sponsored by several Catalan institutions such as the Generalitat de Catalunya, who have contributed to the project with a research grant (ref. FI 96/6.008 PG), the Ajuntament de Lleida, who have financed the project Acalex, the former version of LVBAC, and the University of Lleida, who have sponsored our research group from 1996 to 1998. The members of the research group come from Computational Linguistics and Artificial Intelligence, as well as Foreign Language Teaching at university level and they specialise in Catalan and English.

[2] The research group has already done some research on Catalan verbal morphology in J. Tió and F. Manyà (1993).

mind come in handy. What they advocate is some kind of post-symbolic semantics where things are thought to mean what they are. Concepts, whether as mental concepts in cognitive science or as representations of reality as we find them in the so-called language-independent ontologies used in expert systems, are to be processed not as symbols, but so to speak at their iconic face values. Language, on entering the realm of artificial intelligence and cognitive science, sheds meaning and is reduced to a formal calculus where the correctness of expressions is decided solely on syntactic grounds. Computation means algorithmic permutation of strings of uninterpreted symbols usually called concepts. Is artificial intelligence or cognitive science really the place to turn to in our quest for the phenomenon of meaning? "For the purpose of the present investigation, word meanings are just concepts", we are told by Jerry Fodor in his newest book (Fodor 98, p. 2). This is what he has to say about the meaning of (the concept of) *doorknob* (Fodor's favourite concept): "What has to be innately given to get us locked to *doorknobhood* is whatever mechanisms are required for doorknobs to strike us as such". (Fodor 98, p. 142) Our concept of doorknob, it seems, is somehow metaphysically 'locked' (metaphors abound when talking about things we find difficult to grasp) to doorknobs as they occur in real life. And since doorknobs come in different sizes, colours and shapes, we might add, the most real doorknob will be the one ideal doorknob consisting of nothing but its essential features. Post-symbolic philosophy of mind takes us back all the way to Plato, where human souls are endowed with innate memories (concepts) of the ideal doorknob or with an innate mechanism that will create the concept of doorknob when confronted with one. Is this where linguists should look for meanings?

The mainstream philosophy of mind is firmly rooted in the tradition of analytic philosophy. For Kevin Mulligan, in a recent review in Times Literary Supplement, there is an important 'divide' between 'analytic' and 'continental' philosophy: "Analytic philosophers spend their time [...] elucidating some proposition, analyzing and describing, drawing distinctions and constructing theories. Continental philosophers spend their time creating concepts and conceptual poetry, subverting, suspecting, unmasking, decoding, deconstructing and intuiting [...] entities that are rarely as manageable as some particular thesis or theory [...]" (Mulligan 1998, p.6). It is no surprise that analytic philosophers keep easy company with cognitive scientists, when it comes to dig down to the bottom of meaning, or better, language understanding. The other branch of philosophy (whom I would prefer to call hermeneutical rather than continental) find their home in (post-)structuralism and, more generally, in the social sciences. For them, meaning is not a mental concept that eventually can be identified with specific brain cells ('neurons') and their synapses but rather social fabric that can be negotiated by the language community at large, and its residue can be found nowhere else but in the textual transactions between the members of this community.

Analytical philosophy may let you choose between the mind as the garden of meanings and metaphysical reality as the quarry of conceptual truth. But this distinction set aside, both the mental and the metaphysical factions agree that concepts are atomistic and strictly monosemous, that they are distinct and that they can be manipulated (without being interpreted) by performing truth-preserving syntactic operations on them. Hermeneutical philosophers think meaning is symbolic, fuzzy, text-intrinsic, context-dependent and that understanding and interpreting a text is an action involving intentionality, i.e. an action that cannot be reduced to a sequence of algorithmic computations. Translating a text presupposes text understanding. Otherwise it would be possible to translate a text from language A into language B without knowing either of them. It is indeed not possible, not even with the best dictionaries.

The core issue of translation is meaning. Therefore linguists in quest of meaning should perhaps turn to their even poorer relatives, the translators. For the only kind of meaning linguists, I think, should be interested in is meaning that can be conveyed verbally. Language, after all, is also a social

phenomenon. Verbal communication is meaningful. Linguistics is not concerned with the mysteries of minds. It is not possible to know (and it makes, for linguists, no sense to strive to know) how the participants in verbal communication understand the words, sentences, and texts they say or hear. Understanding is a psychological, a mental phenomenon. As a social phenomenon, language manifests itself in texts, and only there. Translators trade in texts, and they promise that they can paraphrase a text in a different language so that the paraphrase will mean almost the same as the original text. Of course, to carry out their task, they have to understand the text. This means that they interpret it before they paraphrase it in the target language. Which tells us that for the translation of unrestricted general language, computation, i.e. the permutation of uninterpreted symbols, won't do. This is really not such a new story. Alan K. Melby has, in his book *The Possibility of Language*, given us a thorough account why machine translation based on conceptual ontologies cannot work. But the machine translation community has paid little attention. To engineers brought up with a world view based on metaphysical realism, natural language is but a contaminated variety of formal calculi. To them, the promise of machine translation appears to have lost little of its charm.

After all, it was the discontent with the way meaning is tackled in the classical dictionary, monolingual or bilingual, that made the cognitive science, the artificial intelligence and the machine translation communities search for alternatives. In this study on lexical meaning in translation, I will reassess the bilingual dictionary approach, and I will show why the conceptual ontology cannot be considered as a viable alternative. Starting with German *Trauer* (usually translated into English as *sorrow* and *grief*), I will argue for the corpus linguistics approach in the field of multilingual lexical semantics and in computer-aided translation. I will first show that the semantic information of bilingual dictionaries is not nearly adequate for human translation into a non-native language, and that bilingual lexicons derived from dictionaries are even less adequate for machine translation. Then I shall look at conceptual ontologies. My claim is that any attempt to render them language-independent or language-neutral will impede their usefulness for translation, paradoxical as this may sound. Finally I shall discuss the corpus linguistics view on multilingual lexical semantics and stress the work that still has to be done before we can dream of a translation platform that really helps human translators to deal with unrestricted general language.

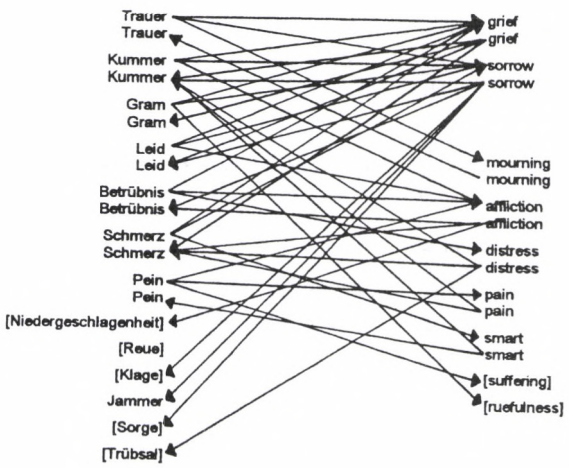
2. The dictionary approach

The bilingual dictionary has always been an indispensible translation tool. It is more valuable for translating into one's own language than into a non-native language, and if it is large enough, it contains all the single words that occur in the text to be translated, even if it falls short of listing the multi-word units, collocations and set phrases. Translators with their knowledge of the foreign language, restricted as it may be, can usually identify such phraseologisms and often will be able to supply a translation equivalent for their native language they cannot find in the dictionary. If they are confronted with a list of equivalents they know to select the one the translation asks for. In equivocal cases, the dictionary sometimes adds semantic and pragmatic categories to facilitate the choice. It is not quite clear how useful this information really is for translating into one's native language. Due to the lack of explicitness, the information provided by the dictionary allows for translating into a foreign language to an even lesser degree. Here amateur translators cannot be expected to supply equivalents of collocations and set phrases missing in the dictionary; neither can they be expected to choose the adequate equivalent from a list if no information on usage conditions is given. Generally this kind of information is deficient.

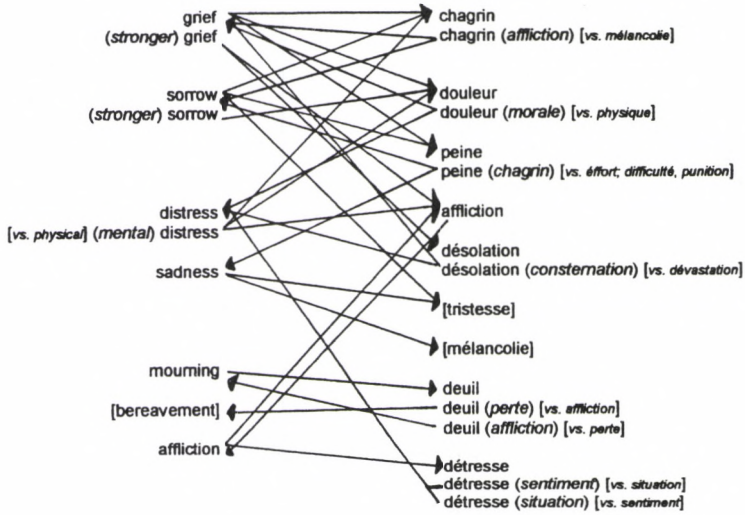
Is it possible to convert bilingual dictionaries into machine translation lexica with the goal of achieving translations comparable to those of human translators? Certainly not; the implicit

knowledge of both source and target language which a human translator has is a necessary precondition for successfully using dictionary information. What is not spelled out in the lexicon is just not available to the computer. There is, in machine translation, no counterpiece to the translator's language competence. But would not a dictionary-based lexicon at least allow for translations comparable to human translations into a foreign language? Can we expect machine translation to select the right equivalent using a dictionary-derived lexicon? It would have to be a dictionary-lexicon containing nothing but unambiguous explicit instructions which would allow us to translate from source language into target language without understanding any one of them. Such a dictionary would resemble the book of instructions used in John Searle's Chinese Room experiment, and is about as realistic.

Let us look at a concrete example. Starting from the with German *Trauer* we look up the equivalents given in a German-English/English-German dictionary (i.e. *sorrow* and *grief),* and then we pursue our search by looking up the German equivalents for these English words. From them German words we again proceed to their English counterparts. We end up with a German and an English word list and arrows denoting equivalence, unidirectional or, as the case be, bidirectional. This process is, in principle, open-ended. We stop it where the equivalents led away from the semantic field with *Trauer* as its core. Our source of information is the 4-volume Langenscheidt Enzyklopädisches Wörterbuch Deutsch-Englisch/Englisch-Deutsch. These are the results of our search:

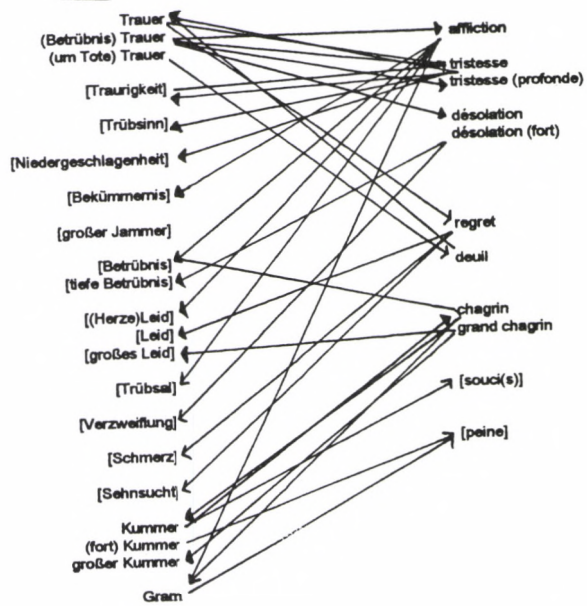


Looking at the result, our first impression is that of chaos. It is surprising how few bidirectional equivalents we find. This may have to do with cultural idiosyncrasies of the two vocabularies of emotions; on the other hand, it may show that different lexicographers, be they German or English, view things differently or just out of context. For a better understanding of our results, we repeat our search for equivalents for English and French. This time we start with the three English words *grief, sorrow* and *mourning* that we found listed as equivalents for German *Trauer.* We also add (in round brackets) the semantic or pragmatic categories we find in the dictionary and where applicable, in square brackets, the categories against which they are juxtaposed. The dictionary used here is the Le Robert&Collins French-English/English-French dictionary.



Leaving aside for a moment the semantic or pragmatic categories, we find a situation very similar to the German-English picture. Unidirectional equivalents outweigh bidirectional ones. Some of the words given seem to be loose ends, like *mélancolie*, *tristesse* or *bereavement*, leading us away from the core of the semantic field indicated by *grief, sorrow, mourning*.

It is now interesting to continue our quest by carrying out the same process for the German and French vocabulary. Again we start with *Trauer*, and, as before, we add the semantic or pragmatic categories. This time we use the Langenscheidt Großwörterbuch Französisch-Deutsch/Deutsch-Französisch.



The picture evolving for French-German equivalences is similar to our previous findings. For German, we find 13 equivalents of 8 French words, and, as was the case with the other two language pairs, all words are somehow connected by equivalence relations.

In all three instances the situation seems to be alike. For the core words of the semantic field of *grief* the dictionaries present a choice of equivalents without clearly indicating how the translator should choose from them. It seems we are left with two hypotheses:

a) the choices given for each word are synonyms, and the translator can select any of them
b) the equivalents non-synonymous; they are used differently, but the lexicographers are unable to provide proper guidelines.

From our linguistic competence we can quickly rule out the first hypothesis. However, if the words given as equivalents are not synonyms, it is the case that the dictionaries fail to give us instructions how to translate these words into the target language.

It would be fascinating to analyze the three tables in view of lexical equivalence. Due to lack of space, I will confine myself to two questions:

a) If I want to translate *Kummer, Gram, Trauer* into English, when do I choose *grief,* when do I choose *sorrow*?
b) If I want to translate *grief* or *sorrow* into German, when do I choose *Kummer,* when *Gram,* when *Trauer*?

Native speakers of English tell me that *grief* and *sorrow* are not synonymous, and as a native speaker of German I know that *Kummer, Gram* and *Trauer* mean different things.

Sometimes bilingual dictionaries give us subtle indications. In table three (with the French-English equivalents) we find, certain expressions in brackets (e.g. *affliction, morale, chagrin*) which refer to one particular sense of the source language word. Thus there are, according to Le Robert & Collins, four senses of *peine*, characterized by their respective categories *chagrin, effort, difficulté, punition,* and their respective equivalents a) *sorrow, sadness*, b) *effort, trouble*, c) *difficulty*, d) *punishment, penalty*. The status of the categories remains unclear, though. Viewed solely under the aspect of the source language, they look rather like hypernyms, but viewed from their target language cognates, they could also count as synonyms. For the French word *deuil* we find two senses of interest to our semantic field, characterized by the categories *perte* and *affliction*, resp., and with the equivalents *bereavement* and *mourning*, *grief*. As instructions they are rather useless. The sentence *Il rient d'avoir plusieurs deuils dans sa famille* cannot be translated by *He had several bereavements in his family lately.* There word to choose here is *losses*, a word not given in the entry.

The situation we find in the Langenscheidt German/English and English/German dictionary is even more unsatisfactory. For grief, two senses we given. 1. *Gram, Kummer, Leid, Schmerz*, and 2. *Unglück, Katastrophe, Fehlschlag*. No indications are given when to use which equivalent. For *sorrow*, we find three senses. 1. *Sorge, Kummer, Leid, 2. Reue*, 3. *Betrübnis, Klage, Jammer*. For all three senses, it seems, a list of synonyms is added: SYN. *anguish, grief, regret, use*. Who is expected to profit from such a list? It does not help with selecting the proper equivalent. It is unpardonable that the most obvious equivalent both for *grief* and *sorrow*, namely *Trauer*, is not listed. For if we look under Trauer, we find four senses, three of which we can disregard for our analysis [i.e. *2.* (Trauern) *mourning*. 3. (Trauerkleidung) *mourning* 4. (Trauerzeit) *(period of) mourning*]. But the first sense given, indicated by the expression (*Schmerz, Kummer*) [hypernyms, synonyms?] list just *grief, sorrow*. Similar indications of unclear status we find for the second and third sense of *Kummer* (but not for the first): 1. *grief, sorrow, heartache*, (stärker) *affliction*, 2. (Sorge, Verdruß) *trouble, worry*, (stärker) *distress*, 3. (Bedauern) *sorrow, regret*. Finally *Gram*: Here we find the category 'lit.' telling us that the German word is used only in serious literature. Is

this relevant for translating it into English? Of the two senses the first is not marked: 1. *grief, sorrow, ruefulness* (lit.); *dolor, dolour* (poet.). For the second sense we find: 2. (Trauer) *sadness, melancholy*. Why *sadness* and *melancholy* are associated with *Trauer*, whereas *grief* and *sorrow* apparently do not fit into this category, will always remain a mystery. None of the information given in this dictionary can be interpreted as an instruction for selecting the proper equivalent.

To come back to the English words in our focus, to *sorrow* and *grief*, we perhaps should pay heed to the admonitions of our foreign language teachers, and take refuge to the monolingual dictionaries when we are left in doubt by the bilingual ones. In our choice between *grief* and *sorrow* as equivalents for *Trauer, Kummer, Gram* we turn to two dictionaries, the 1987 COBUILD and the 1998 New Oxford Dictionary of English (NODE). There we find the following entries:

grief
COBUILD
- grief is extreme sadness
- a grief is something unpleasant that happens which causes someone great sadness or unhappiness (...was a grief to s.o.)

NODE
- deep or intense sorrow, especially caused by someone's death

sorrow
COBUILD
- sorrow is a feeling of deep sadness or regret, or caused for example by the death of someone you love or because of your sympathy for the sufferings of someone else
- a sorrow is an event or situation that causes deep sadness (after all these sorrow and trials)

NODE
- a feeling of deep distress caused by loss, disappointment, or other misfortune suffered by oneself or others
- an event or circumstance that causes such a feeling (... it was a great sorrow to her...)

It seems rather obvious that, according to both dictionaries, the meanings of both words are very similar. We find no indication of a difference that would tell us in which situation one should be used rather than the other. According to NODE, *deep distress* is *sorrow*, and *deep or intense sorrow* is *grief*. According to COBUILD, *sorrow* is *deep sadness* or *regret*, while *grief* is *extreme sadness*. We do not learn from COBUILD how the feeling of grief is caused, but this is probably just an oversight. The most stereotypical cause for both *sorrow* and *grief* seems to be *someone's death*. Apparently *sorrow* can also be suffered as consequence of a *disappointment* or of a *misfortune* that oneself has experienced. But other people's *misfortune* or *sufferings* can count as well, as long as you sympathize with them. We are not told if this is also true for *grief*. The information gathered from both dictionaries, be it by itself or be it in conjunction with the information given in the Langenscheidt dictionaries, cannot be read as an instruction when to translate *Kummer*, *Gram* and *Trauer* into *grief* and when into *sorrow* (disregarding other alternatives).

Thus it seems that neither bilingual dictionaries by themselves or a combination of bilingual and monolingual dictionaries are sufficient when it comes to translate a text into a language which the translator does not speak quite well. It is also obvious that these dictionaries, either alone or in combination, cannot be sufficient as the lexical basis for machine translation, no matter how they are processed. Unlike human beings, a computer does not understand what a text, a sentence means, neither in the source language nor in the target language. The lexical information the computer

needs to translate from language A into language B is equivalent to what a human translator who speaks neither language A nor language B would need. Dictionaries are designed for people with at least a basic knowledge of the language(s) covered. It is not conceivable how lexicons suitable for machine translation could be derived from them.

3. The conceptual ontology approach

In cognitive science, and hence in classical artificial intelligence, including machine translation, concepts seem to occupy the place words have in classical linguistics. Words and concepts somehow seem to correlate. The "somehow" is where authors differ. So we can read (cf. Teubert 98):

- concepts are expressed in words in a natural language
- concepts represent the abstract meanings of words
- concepts represent word meanings
- word meanings are just concepts.

Concepts are thought to be pure, whereas words are thought to be contaminated by noise emanating from the idiosyncrasies and contingencies of natural languages. Concerning the fabric of concepts, we find a diversity of views; all of which can be reduced to one of three main conceptions:

- Concepts can be the result of an explicit agreement between experts. They are, in principle, language independent or, at least, language neutral. This is the view we find in terminology. The concept 'alphabetic character' is what the relevant ISO committee has defined.
- Concepts, at least primitive concepts in Anna Wierzbicka's sense, are innate, features of the mind/brain, so to speak, while they may be occasioned by the pertaining experiences. Complex concepts, in this view, are composed of primitive concepts. This view was endorsed by the early (but not the recent) Jerry Fodor, when he was still a Radical Nativist (in *The Language of Thought,* Fodor 75) and is still the standard paradigm for Stephen Pinker (e.g. in *The Language Instinct*). For him, concepts are the words of a universal mentalese (Pinker 94).
- Concepts exist independently of people's minds. Concepts are the true meanings of things (regardless of what we believe). They are the pure essence, the idea underlying the material. That is the Platonic view. Augustine also subscribed to it, and today, it is the view of the metaphysical realists dominating the American philosophy of language, notably Hilary Putnam (Putnam 81).

In one aspect, all three views concur: Concepts are language-independent or language-neutral. This is (at least to some extent) true for terminologies and for that part of the vocabulary where words share certain features with terms (like words denoting natural kinds). Terms are said to have no meanings; they have definitions instead. These definitions are stored in certain authoritative texts (like glossaries or term banks), and what is not contained in the definition, is not essential. Terms are concepts in the sense that the concept is co-extensive with the term definition. Still we should bear in mind that definitions are given in a natural language, and even where we find a formula instead (like 'H_20' for water) this formula will have been explained somewhere in natural language (so that we might know what it means). Nevertheless, since for terms concepts are co-extensive with their definitions, these concepts can be called language-independent or language-neutral to the extent that the definitions are fully translatable. If we could find an ontology of concepts denoting emotions, if in this ontology the concepts are co-extensive with definitions, and if these definitions are translatable, then such an ontology would be the answer to our problem of translating *Trauer* into English and *grief* and *sorrow* into German. There is no such list, and, in any case, one problem remains: Can we reduce the meanings of words describing emotions to definitions, i.e. definitions

that capture all essential semantic features so that what is not contained in the definition is not essential?

While conceptual ontologies are ideally language-independent or language-neutral, existing ontologies have their concepts including their definitions spelled out in natural language. Most existing ontologies do not attempt to cover everything that can be said or thought about the world in general. Usually they represent a limited domain, some particular field of expert knowledge. There are two exceptions, however, the ontologies CYC and WordNet; they both have a rather general scope including concepts, or words, denoting emotions. Both ontologies use English in their definitions, and both correlate the definitions to English expressions. While WordNet never claims to be anything else than a very large and very elaborately connected electronic thesaurus of the English language, it is nowhere explicitly stated which status the expressions used in the definitions (hypernyms, hyponyms, synonyms etc.) have: Are they just English words with all their usual fuzziness of meanings or are they concepts meaning nothing else than what their definitions say? In other words: Are the definitions expressions of a formal language or are they natural language expressions? For the CYC ontology, there is a representation language CYCL (a "First Order Predicate Calculus with some second order features") and there is an algorithm that will translate from English to CYCL.

The concepts correlated with the words *sorrow* and *grief* and their definitions given in CYCL are not accessible on the Internet. (For other concepts, like *embarrassment, shame* or *guilt,* see: http://www.cyc.com/cyc-2-1/vocab/emotion-vocab.html). In our analysis, we look at WordNet and focus there on *sorrow*. We are interested in information telling us under which conditions *sorrow* has to be translated into *Trauer*, or *Kummer*, or *Gram*, respectively. In WordNet, we find the following information on *sorrow* (http://www.cogsci.princeton.edu/cgi-bin/web.html):

4 senses of sorrow

sense 1
sorrow – (an emotion of great sadness associated with loss or bereavement; "he tried to express his sorrow at her loss")
⇒ sadness, unhappiness – (emotion experienced when not in a state of well-being)

sense 2
sorrow, regret, ruefulness – (sadness associated with some wrong done or some disappointment, "he drank to drown his sorrows")
⇒ sadness, unhappiness – (emotions experienced when not in an state of well-being)

sense 3
grief, sorrow – (something that causes great unhappiness; "her death was a great grief to John")
⇒ negative stimulus – (a stimulus with undesirable consequences)

sense 4
sadness, sorrow, sorrowfulness – (the state of being sad; "she tired of his perpetual sadness")
⇒ unhappiness – (state characterized by emotions ranging from mild discontentment to deep grief)

The words occurring with sorrow, i.e. *regret, ruefuless, grief, sadness, sorrowfulness* are seen as synonyms in relation to the respective senses. In the scope of the definitions given for these senses, they can be substituted for each other. Comparing the four senses with the two senses of NODE and

COBULD (which are for our purposes interchangeable), we find that WordNet sense 1, 2 and 4 are conflated into sense 1 of NODE/COBULD. WordNet sense 3 correlates with sense 2 of NODE/COBULD. The breakdown of NODE/COBULD sense 1 into three senses shows in an nutshell the issues at stake with this 'Thespian' part of the vocabulary (Sinclair 96). Words like *sorrow* may be said to have some kind of prototypical meaning (Lakoff 99), but lexicographers seem incapable of finding a plausible definition that captures all the essential features and nothing but them. As WordNet shows us, there are some citations where *sorrow* can be substituted by *regret* (really without changing the meaning?), though probably not by *ruefulness*. And there are other citations where *sorrow* can be replaced by *sadness* (or the rather rare *sorrowfulness),* but not by *regret*. If this were to work on a convincing scale, we could solve the issue of semantic fuzziness by explaining it in terms of family resemblance. But the WordNet entry shows convincingly that it does not work. For are *grief, woe, affliction, distress, melancholy* not synonyms of *sorrow* in senses 1, 2 or 4? Is WordNet sense 1 of *affliction* "a state of great suffering and distress due to adversity" really not synonymous with sense 1 or 4 of *sorrow*? Neither Fillmore's prototypical meaning nor the family resemblance approach are truly convincing.

The second important issue that can be exemplified by this WordNet entry is the status of the definitions. Are the definitions given in a controlled language where each word is monosemous and strictly defined, or are they given in unrestricted general language? Do words used in the definitions like *sadness, feeling, emotion, state, psychological feature* always have the same meaning (i. e. are they used as terms), or are they just as fuzzy as the words they are defining? The explicit hierarchy of hypernyms goes far beyond what a general language thesaurus offers; it also gives cohyponyms ('coordinate terms'), meronyms, toponyms and even antonyms (for *sorrow* sense 1: *joy, joyousness, joyfulness*), and its overall architecture is certainly that of a typical conceptual ontology. However, if we look closer at the hierarchy of hypernyms, we find that they are identical for senses 1 and 2, but different for 4. The 'synset' of sense 4 consists of *sadness, sorrow, sorrowfulness*, and the synset of the hypernym of senses 1 and 2 is *sadness*, *unhappiness*, and the hypernym of ths synset is *feeling*, while in sense 4 we find *unhappiness*, then classified as *emotional state*, whereas in senses 1 and 2 it is classified as *feeling*. There are more inconsistencies, as with *feeling* and *emotion*. Though *emotion* is defined as 'any strong feeling', we find that *sorrow* (sense 1) is defined as an 'emotion of great sadness', that its hypernym is *sadness*, and that the hypernym of *sadness* is *feeling*.

These inconsistencies take us directly to the third issue at stake. What do the concepts in an ontology stand for? Does the WordNet vocabulary of emotions describe general language words (whose meanings are embodied in the way they are used by the language community) or does it describe the terminology as it has been agreed upon by psychologists? Can a psychologist use the term *unhappiness* to describe what a patient calls her *sorrows*? Are the entities described in WordNet concepts in this sense? Or can they be thought of as language-independent or language-neutral (not in the Princeton WordNet undertaking but perhaps in the European offspring EuroWordNet), because they can be postulated as abstractions composed of universal semantic primitives in the sense used by Anna Wierzbicka? Then each word (or perhaps each sense of a word) belonging to any language would be represented by a concept. Sometimes, if words belonging to one language or to different languages are truly equivalent, they would be represented by the same concept; and in those cases where we do not find a true equivalent in the target language, we would look at the composition of the concept representing the source language word, find the target language word whose concept is closest to the source language concept, and maybe use an appropriate adjective that takes care of the semantic difference: *distress* is not just *Kummer*, but *starker Kummer*.

What, then, is the advantage of the conceptual ontology approach? Let us recall that our goal is to facilitate the translation of unrestricted general language using the computer. Conceptual ontologies aiming to represent language-independent or language-neutral concepts so for have failed to demonstrate their language independence. Anna Wierzbicka's approach, however, seems to assert the postulate of language independence. But how would it be possible to establish the universal primitives it presupposes? In which language would they be defined? If the definitions were in natural language, the problem of circularity could not be avoided. If the definitions were in a controlled language (like an algebraic calculus), how would they relate to the natural language which they are supposed to cover? But, for the sake of the argument, let us assume for the moment that a fully blown language-neutral conceptual ontology (perhaps of the CYC type) exists. Would or could the translation of unrestricted general language profit from it? Taking into account Wierzbicka's model, would we not have to translate a text first into a conceptual representation and then this representation back again into a natural language text? That means doubling the effort. more to the point, would it work anyway? Alan Melby has told the story of artificial intelligence applications and machine translation systems based on conceptual ontologies. It is a tragic story, a story of unavoidable, cogent failure.

4. The corpus linguistics approach

If neither the dictionary/lexicon approach nor the conceptual ontology approach (nor a combination of them) offer much help when it comes to translating texts such as newspaper articles into a foreign language, perhaps we should start looking elsewhere. Whenever we compare a text with its translation we realize that the real issue is not the single word. No human translator translates an unrestricted general language text word by word. Translators interprete a text sentence by sentence, and based on their interpretation, they will formulate target language paraphrases for these sentences. Of course, there will be some words for which there are lexical matches readily available. Regardless of context, a *lion* will always be a *Löwe*, an *elm* will always be an *Ulme*, and a *refrigerator* will always be a *Kühlschrank*. But often the proper translation equivalent is a question of context, of interpreting the whole sentence, perhaps not even just by itself but in the light of all preceding sentences. Whether a *Berg* is a *mountain* or just a *hill*, whether *Marmelade* is *marmelade* or *jam*, a *Straße* is *street* or *road* is determined by the contexts in which these words occur. From the German point of view, *Berg, Marmelade*, and *Straße* are perfectly monosemous words, they are not ambiguous in the sense that they represent two different concepts. They are monosemous, but fuzzy, and their fuzziness does not map with the fuzziness of *hill, mountain, marmelade, jam, street,* or *road*. A *Straße* is a *street* if there is a row of houses on its sides, if it is in a town or a village; if it is in the countryside it is a *road*. The context will tell the translator where the *Straße* is so that she or he can make the right choice. If we analyze the contexts of a given word in an fairly large corpus using quantitative methods we often recognize recurrent context patterns, text segments that tend to come up time and again in the same or similar form, with the same or semantically related words, frequently in the same or related syntagmatic shape. Those text segments often constitute entities like compounds, multi words units, collocations and set phrases, which must be translated as a whole. A compound like *grief work* is stable; there is no *sorrow work* or *grief labour*. But outside of these kinds of stable text segments, it can be quite difficult to determine the relevant context elements indicating whether *grief* is to be translated into *Trauer* or *Kummer* or *Gram*. We may need many hundreds of corpus citations of sentences with *grief* together with their German translation to discover the underlying context patterns, and with words like *grief* we may will find out that the relevant context is much larger than just a simple sentence, that it could be the whole paragraph or even chapter.

For translating words belonging to the "Thespian" part of vocabulary there is but little advantage, it seems, to gain from subjecting them to the conceptual ontology approach. Whether we assign four senses to *sorrow*, as in WordNet, or claim *sorrow* is a conflation of the three concepts *Trauer*, *Kummer* and *Gram*, or whether we just say it has a fuzzy meaning, we always have to analyze the contexts, we always have to come up with the relevant text segments in order to select the appropriate translation equivalents. By isolating them and looking them up in a dictionary (or a conceptual ontology connected with a lexicon) we forego the chance to resolve their inherent fuzziness. With a dictionary listing the complex translation units that I called text segments it would be much simpler to translate. But here are good reasons why there is no such dictionary of translation units. The only way to discover and identify them is in aligned parallel corpora, and there are still not very many around. Since the detection of translation units uses recurrence as default, we would need very large parallel corpora to find the less obvious complex translation units, i.e. those that have not already been listed in bilingual dictionaries. But the real problem is the dynamics of translation units, their instability, flexibility, variability, features related with their size. Neglecting the function words, we often find that some lexical elements of such units can be replaced by others, and that, while it is possible to describe some text segments as relatively stable syntagmas (eg. compounds like *grief work* or collocations like *put to grief*), others are much more variable and unpredictable (like text segments with *grief*, but excluding compounds and collocations), and cannot be reduced to stable syntagmatic patterns. Text segments of this kind simply cannot be listed in a dictionary.

Here is a small sample of sentences with *grief* and *sorrow*, taken from the BNC (and syntactically shortened where possible):

grief

(1) I experienced tears of real grief the first time when my granny died.
(2) My one real consolation in the face of almost any betrayal or grief is that I am without doubt an extremely talented writer, come danger or candlestump.
(3) Grief gave way to a guilt that gnawed at him.
(4) In close communities like mines, the workers have been able to express both grief and anger openly.
(5) She had heard the news with regret, with sadness, but hardly with grief.
(6) So much of life, its relationships and its creative opportunities are damaged or lost in the course of addictive disease that grief is universally a major factor in early recovery.
(7) The grief had none of the sanctions of legitimacy.
(8) You mean stricken with remorse or grief?

sorrow

(9) Miller poignantly captures the variety and conflict of feelings that followed the war - the mixture of relief, sorrow, dissatisfaction, and what came to be called survival guilt.
(10) A magic harp music made its listeners forget sorrow.
(11) So that was their secret sorrow.
(12) His fervent soul was full of sorrow for the world and its sinfulness.
(13) For a parent to hide sorrow or even anger denies the child's right to have their own feelings.
(14) It comes from knowing that no matter how intense a pain might be or our sorrow or our anxiousness, that Jesus Christ is the ultimate victor.
(15) I can only say that the parting was such sweet sorrow.
(16) No, we were expressing our sorrow, Graham, I assure you.

Our interest in analyzing these citations is focused on two issues. First we would like to find out if there are observable differences in the way *sorrow* and *grief* are used. Such a distinction would be valuable for translations into English. The other question concerns English as source language. Translating these sentences into German, when would we translate *sorrow* or *grief* as *Trauer*, when as *Kummer*, when as *Gram*? Do the contexts tell us the answers?

Native English speakers agree that in most of our sentences *grief* and *sorrow* cannot be substituted for each other. In sentence (12), however, such a substitution (*grief* for *sorrow*) seems to be possible: in the context of a *fervent soul* the intensity associated with *grief* rather than with *sorrow* would have been the standard expectation. This is why in (8) *sorrow* would not do; but there is a second reason: *stricken with grief* has a collocational quality. *Sorrow* is described as a longer lasting, deeper, but less intense, more introverted than extroverted, more individual than communal, more voluntary than ritualized, more personal than culturally mediated feeling. This is why bereavement causes rather *grief* than *sorrow*. And this also explains why in (10) the harp cannot make its listeners forget *grief*, why in (3) *grief*, but not *sorrow* can give way to guilt, why in (5) *sorrow* cannot replace *grief*, and finally why the Great War eventually led to a feeling of *sorrow* rather than of *grief*.

But how can we automatically disambiguate *sorrow*-contexts from *grief*-contexts, a task necessary for finding the proper equivalent of *Trauer*, *Kummer*, *Gram*? Often it seems to be impossible. A first step could be to look at the cohyponyms of *grief* and *sorrow* we find in the context. In (4), there is the intense, sudden feeling of *anger* joined with grief. In (9), *sorrow* co-occurs with *relief*, *dissatisfaction* and *guilt*, longer lasting, less intense feelings. Then there are juxtapositions like in (3): only *grief* can give way to *guilt*, while *sorrow* could easily co-exist with guilt. Similarly in (5) *grief*, but not sorrow, can be contrasted with *regret* and *sadness*. Still we are puzzled by (12), where the intense *fervent* co-occurs with *sorrow*, also by (13), where *sorrow* (and not *grief*, like in (4)), is conjoined with *anger*, and also by (16), where the more introverted *sorrow* finds outward *expression*. In (14), we suspect that *sorrow* is the preferred term in religious speech, and that neither the attribute *intense* nor the cohyponym *anxiousness* are sufficient to change it to *grief*.

A large parallel corpus of German texts with their English translations and English texts with their German translations would yield enough citations of *grief* and *sorrow* with their German equivalents, and a quantitative analysis of the contexts could give us typical patterns of *grief*-contexts versus *sorrow*-contexts. If we add a symbolic or categorial analysis of the context words in question, building up what is called synsets in WordNet, not from dictionary information but from corpus data as described above, our context patterns would consist not just of fixed lexical items but also of categorial elements, making them much more manageable. The processing of text segments in large parallel corpora would thus provide us with contextual patterns for words like *sorrow* and *grief*; and if we now have to translate a new German sentence with *Kummer*, *Gram*, or *Trauer*, the computer could compare the text segment in which this word occurs with the contextual patterns of *sorrow* and *grief*. It then could indicate the closest match. There still might be counter-intuitive cases and cases where the context leaves us helpless. But even a success rate of more than 50% would be a huge improvement compared with the dictionary or the conceptual ontology approaches where no recommendations are offered.

The second question seems to be easier to answer. We can identify those sentences where *sorrow* or *grief* would be translated into *Trauer* or *Kummer* or *Gram*, resp. *Trauer* is the proper equivalent for sentences (5), (7), (8), (9) and (16). In (5), *grief* is opposed to *regret* and *sadness*, which excludes *Kummer* and *Gram* and leaves only *Trauer*. In (7), *sanctions* and *legitimacy* points to the communal aspects of *grief*, which go together only with *Trauer*, i.e. *bereavement*. Sentence (8) presents a

collocation: *stricken with grief* is equivalent to German *von Trauer geschlagen*. It could not be *Gram* here: *Gram* does not set in suddenly. In (9), *survival* points to death and bereavement and thus leaves us only with *Trauer*. Finally, in (16), *expressing* again alludes to communal aspects and thus calls for *Trauer*. There are some clear cases of *Kummer*: (11), (13), (14) and (15). Actually in (11) and (15) we find *sorrow* embedded in collocations: *secret sorrow*, *such sweet sorrow* (a Shakespeare quote from Romeo and Juliet), equivalent with the German collocations *geheimer Kummer* and *süßer Kummer* (*süße Wehmut* might even be a better translation). In (13), *sorrow* occurs with *anger*, both describing the parents' feelings about their offspring's misdeeds. Children so young that they still have to learn to deal with their feelings can feel only *Kummer* and not *Gram*, i.e. a feeling of high intensity, suddenly setting in, but not of long duration and not of depth. In (14) it is the text type of the religious sermon which tells us that *Kummer* is the right German equivalent. *Pain*, *sorrow* and *anxiousness* will be redeemed by external circumstances, something that could not happen with *Gram*. Good candidates for *Gram* are sentences (2), (6) and (12). In (2), *grief* has nothing to do with loss, and the writer, in all irony, is talking of serious situations, not of predominantly feminine feelings like *Kummer*. In the context of (6), *grief* is not caused by bereavement but by acknowledging an irreversable damage to one's social life is lost beyond recovery, causing a lasting feeling of deep sadness. Such a feeling is also described by *sorrow* in (12); and even though *sinfulness* evokes religious speech, there is no redemption at hand: the sorrow is there to stay. Thus *Kummer* is ruled out.

In twelve of the 16 sentences, we found enough context data to choose the German equivalent, and often we could point to specific lexical elements in the context that determined our choice. In other cases (such as (12)) we can rely on the rule of thumb that ugly old men feel *Gram* while beautiful young girls experience *Kummer*. But what about (1), (3), (4), (10)? At first glance, (1) seems to be a straightforward case of *Trauer*: granny has died. But the context does not tell us how old the narrator was then. In German, children younger than perhaps ten years are not thought of being able to feel *echte Trauer*; they do not know yet what bereavement really means. Instead, they are thought to feel *Kummer*. In cases like these, the context scope has to be enlarged. Likewise in (3) the context contains no hint either to the cause of *grief* or to the person experiencing it. Finally, in (4) we encounter an occurrence of *grief* where none of our favourite three equivalents work. I would suggest a trendy word like *Betroffenheit*, which goes well together with *Zorn* (for *anger*).

A quantitative-categorial analysis of a sufficiently large parallel corpus, where we would find hundreds of citations each for *Kummer*, *Gram* and *Trauer*, this time for translating from English into German, would give us the context patterns, consisting again of fixed lexical items and of semantic categories or synsets, that would tell us how to translate new sentences with the words *sorrow* or *grief*. Again, we problably cannot expect a success rate much higher than 50%. But eventually the outcome can be improved once robust surface parsing is available. Then lexical and syntactic information can be merged into frames, scripts, and lastly into propositions. Contextual patterns based on proportional data can be expected to be by far more powerful than patterns using only lexical information. In the end, we can hope, the computer will be able to give adequate recommendations in the majority of cases while indicating when insufficient data require a human decision.

5. Conclusion

It seems that the problem of translating unrestricted general language with computational support can best be tackled by an approach based on corpus linguistics. As long as our parallel corpora are large enough, our quantitative and categorial tools do their jobs well, and robust surface parsing is available, we can process context data into patterns that can be matched with the text segments of

new texts to be translated. The parallel corpus is our repository of translation units and their equivalents. If we succeed in generalizing these translation units and their equivalents into patterns we will be able to match them with new units or text segments. After all, everything that is not a neologism has been said before and a large enough parallel corpus should contain it. To translate the remaining two percent of neologisms and the few cases where semantic fuzziness cannot be resolved by a categorial analysis of context data human intervention is indispensible.

The corpus linguistics approach does not exclude the dictionary and conceptual ontology approaches: rather it complements them. The lexical information of mono- and bilingual dictionaries is the base on which contextual information is assembled. Conceptual ontology provides us with a strategy to deal with the categorial or symbolic aspects that make it possible to conflate individual contexts into context patterns. It provides a model for describing the hierarchical and horizontal relationships between words. Finally frame syntax helps to transform quantitative information into proportional information.

What, then, is so unique about the corpus linguistics approach? For corpus linguistics, the single word is not the most essential unit of meaning. Rather it is the word as it is used in texts, it is the word as it occurs in a context or a text segment. As everyone knows, texts are not translated word by word but generally by larger units. The core vocabulary of unrestricted general language is inherently fuzzy. It is not the isolated word but the word within its context for which we have to find the proper translation equivalent. What words mean cannot be grasped by mapping them onto a conceptual ontology. Lexical meaning is not constituted by „locking" them to items in the real world. Lexical meaning exists solely in the universe of discourse as the complex web tying words to other words via the contexts in which they occur and tying words within their contexts to all other occurrences of these words in all preceding texts. Precisely this course of action offers a new approach to the issue of translation with the help of computers.

6. List of references

Fodor, Jerry A. (98): *Concepts. Where Cognitive Science Went Wrong*. Oxford: Clarendon Press

Melby, Alan K. (95): *The Possibility of Language. A Discussion of the Nature of Language with Implications for Human and Machine Translation*. Amsterdam: John Benjamins.

Mulligan, Kevin (98): *The Great Divide*. In: Times Literary Supplement, June 26, p. 6-8.

Langenscheidt Enzyklopädisches Wörterbuch Deutsch-Englisch / Englisch-Deutsch. Berlin 1978

Langenscheidt Großwörterbuch Deutsch-Französisch / Französisch-Deutsch. Berlin $1979^{6}/1979^{1}$

Le Robert & Collins Dictionnaire Français-Anglais / Anglais-Français. London: Collins 1995^{4}

Collins COBUILD English Language Dictionary. London: Collins 1987

The New Oxford Dictionary of English. Oxford: Clarendon Press 1998^{1}

Teubert, Wolfgang (97): *Translation: Ontologies for Multilingual Applications?* In: Proceedings of the Workshop "Corpus Use and Learning to Translate". Bertinori/Forlí. Available at: http://www.sslmit.unibo.it/cult.htm

Fodor, Jerry A. (75): *The Language of Thought*. New York: Crowell

Pinker, Steven (94): *The Language Instinct. How the Mind Creates Language*. New York: William Morrow

Wierzbicka, Anna (96): *Semantics. Primes and Universals*. Oxford: Oxford University Press.

Putnam, Hilary (81): *Reason, Truth and History*. Cambridge: Cambridge University Press

Sinclair, John (96): The Empty Lexicon. In: International Journal of Corpus Linguistics I/I, p. 99-120

MoBiGloss: A Virtual Dictionary System on the Internet

LÁSZLÓ TIHANYI

Abstract

More and more dictionaries and glossaries are published on the web. There are people, mainly translators, whose daily activity is collecting glossaries from the Internet, and maintaining their collection. There are several problems with this activity. First, web sites frequently have statements that prohibit downloading. Second, downloading is a time-consuming, expensive and frequently exhausting job. Third, results become obsolete immediately, since Internet glossaries are improving very fast. A better tool is needed. A tool that solves the above-mentioned problems and also ensures that all the world's available Internet dictionaries are looked up, so collecting has no sense any more. Something that spares time of visiting multiple sites (or checking variouis downloaded dictionaries) by searching all of them at once. Well-known search engines are insufficient for this task. Although they use different techniques to index sites, they do not recognize dictionaries and cannot differentiate headwords from content. They generally finish indexing after title or, in better cases, the first paragraph of the document. This paper inrtends to give an overview on both of the work how we have created the *MoBiGloss* dictionary search engine, and a description of the user interface.

1. The MoBiGloss project

1.1 Collecting addresses of dictionary websites

There are many dictionary collections on the web. First, we collected addresses of the known collect-sites. It has been an interesting work to compare their contents to each other. Some of them have been proven simple copies of others, further sites have been modified or improved versions of the original sources. A very valuable site we have found is a translator's forum on the net. After a

dozen of collected sites, we concluded that we had collected the majority of the available glossaries. Now, we are working on another tool (a robot) that would discover dictionaries by visiting the sites and analysing their contents.

1.2 Continuous checking of accessibility of the site and availability of the content

Internet connection is not always reliable. Failures might be casued by several reasons. If a site continuously does not reply, it is very likely to be deleted from our database. It is an administrative task to keep this information always correct. It helps to prevent unwanted lookups in dictionaries that do not exist on the net any more.

1.3 Extraction tool for headwords

To get the relevant index items (headwords) we had to analyze the source of the dictionaries in question. Then we defined an algorithm for each site how to get the list f headwords from them. It has been a very interesting work, from the lexicographer's point of view. It is well-known that there exists a proposal for encoding SGML-dictionaries (*Text Encoding Initiatives*), but there are no guidelines for HTML-based dictionaries. Anyone can imagine the result: it is hard to find even two dictionaries that are identical concerning their internal syntax. Everybody shall see how the situation would change when XML (with complete freedom of tag naming and structure) will be commonly used. We would happily support volunteer dictionary writers by giving them hints how dictionary entries should be edited. On the other hand, only dictionaries the publishers of which permit studying their dictionary structure are to be presented by our meta-dictionary tool.

1.4 Indexing the headword database

For indexing the lists of entries, we use the index server of Microsoft (*Microsoft Index Server*). This tool is comfortable for our task, since we never have to reindex anything, just update the headword-lists, so everything can be done automatically.

1.5 Maintenance

Maintenance is not a lexicographic issue. This work with the given tools and background could be set up in a relatively short time. But continuously keeping a public service like this alive generates costs. Therefore, the question of maintenance of a meta-database is not trivial.

2. The MoBiDictionary Internet interface

MoBiGloss Internet glossaries —similarly to other dictionaries— published by MorphoLogic are available via *MoBiDictionary*, our Internet dictionary interface that can be reached on the following address: http://www.mobidictionary.com. In fact, it is a CGI-program written in C, and can communicate to various dictionary servers. So far, we have implemented two servers: the *MoBiDic Server* and the *MoBiGloss Server*. The *MoBiDic Server* is an SGML-based dictionary server presently with more than fourty mono-, bi- and multilingual dictionaries. The *MoBiGloss Server* currently has access to about 400 Internet glossaries. Naturally, these dictionaries can be opened and closed at the user side, as a personalised operation. Further dictionary servers communicating with on-line Internet dictionaries are soon to be added to the service. In *Figure 1* we can see the user interface of *MoBiDictionary* with a word looked up and found in several dictionaries, and *Figure 2* shows hits provided by *MoBiGloss*. The upper frame with the logo contains the input window. Words, phrases or occurrences of words in phrases can be searched with the help of *MoBiDictionary*.

- There are two search options: simple search and phrase search (all the words of a multi-word expression are keywords).
- The source language and the target language of thw query can be selected, but there is a special option, the so-called 'any language' option for unknown input language or multi-language output.
- The list of available dictionaries can also be customized. The user can open or close any of the available sources according to his/her needs.
- There is a set of technical parameters that can be changed using the Preferences button.
- The user language also can be chosen (for the time being, there are only two user languages: English and Hungarian).

After entering the input word, the program fills in the frames below. The left frame in the middle with the pair of flags contains the hits that tells you which dictionary gives hits concerning the word looked up. The hit needs not be identical to the search string: in case of simple search, because of the stemming function, and in case of a phrase search, because of the fact that all the words in a multi-word expression are keywords. Clicking on one of the hits of the left upper frame results in filling all the other frames with the information connecting the certain occurrence of the hit. The lower left frame contains the alphabetical list of neighbouring headwords. The length of this list can be configures by the users, but limited because of safety reasons.

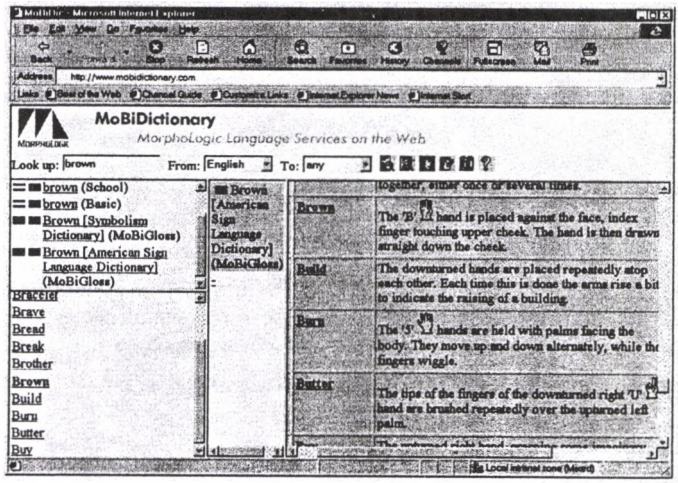


Figure 1. MoBiDictionary

The frame in the middle contains (mostly other language) equivalents of the word. We have introduced this auxiliary frame because we wanted to give a simple list of meanings for the word that has been looked up. Lexicographers frequently try to make the entries of paper dictionaries as short as possible to reduce publishing costs. This is done by the use of parentheses, slash, etc. characters that make dictionary brief (and give an impressive outlook) but which is completely useless in case of electronic dictionaries. In the latter case, we do not have space problems when we visualise an entry, moreover we want to reach the result in one click. (This window is empty in

Figure 1 since in monolingual glossaries words have explanations rather then equivalents). The rightmost frame shows the dictionary entry itself. In case of Internet dictionaries, this information has been loaded from the original website. Dictionary lookup in *MoBiDictionary* is supported by morphological lemmatizer modules (for all the languages supported by the *Humor* morphological engine of MorphoLogic).

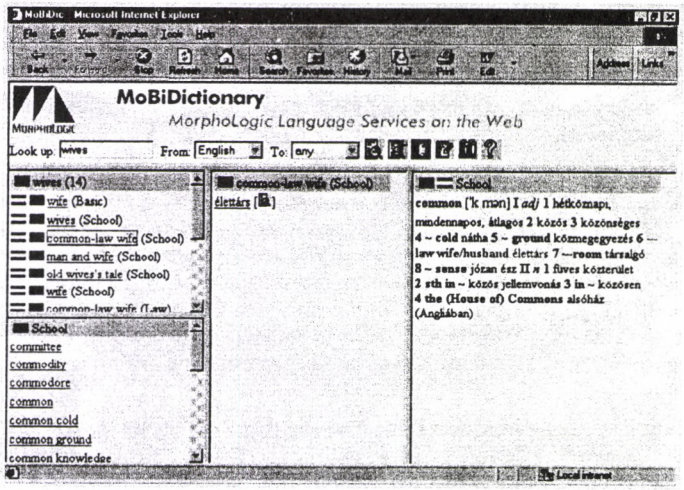


Figure 2. MoBiGloss

3. The Content of MoBiGloss

Internet dictionaries vary in content, size, format and language. From the project's point of view only the last two are important issues, but let us have some comments on their contents and languages. We categorized the dictionaries and got the following statistics.

The list of glossaries indexed by *MoBiGloss* (in order of magnitude):

Technical	90
Miscellaneous	37
Medical	31
Finance	27
Computer	24
General	19
Sport	15
Computer; Internet	15
Biology	12
General; Slang	8
Literature	7
Maritime	6
Environment	5

Cooking	5
Chemistry	5
Transport	4
Legal	4
Economy	4
Insurance	3
Geography	3
Weather	2
Water	2
Trade	2
Sailing	2
Marketing	2
Management	2
Linguistics	2
Forestry	2
Translation	1
Recycling	1
Religion	1
Politics	1
Philosophy	1
Math	1
Literature	1
Geology	1
Food	1

Some remarks on the above list:

- The technical glossaries are generally written to cover terminology and jargon of some specific field.
- The intention of publishing financial glossaries, as we could judge it, is rather to increase the number of visitors on the home page of a specific bank or insurance agency.
- Medical glossaries may also have this reason in background, but these are rather initiated by selfless public service organizations.
- Sport, cooking, slang, acronym dictionaries are mainly published by private persons, as a hobby.
- General category is relatively small. This is the segment where the demand is much stronger then the support. The effort to create these dictionaries is still bigger than the above interests.

The present coverage of the dictionaries is still inconsistent. Some field e.g. computer science is overrepresented, while others are neglected. We hope that eventual gaps will be smoothed. Altogether: MorphoLogic believes that this is the best way of terminological data publication. Everybody can build and update his/her additional dictionary, at any time. The only problem is to supply appropriate search tools that communicates with them and present their content to the users. Paper dictionaries never give sufficient solutions. If they were specialised, then they did not cover enough. Even if they were big, then they would become outdated too soon. Internet must be a solution to this problem.

Some comments on the languages of the *MoBiGloss* dictionaries. English ultimately dominates. It is very interesting to see that besides the monolingual English glossaries (that explain the meaning of the terminology) the rest are mainly multilingual lists. We practically cannot find real bilingual ones.

4. Further plans

4.1 Extension of list with on-line dictionaries on the web

There are many on-line dictionaries on the net that cannot be downloaded, but can be looked up by filling a query form. We also would like to support access to these dictionaries from *MoBiGloss*. For static glossaries, we always know that the query has hits in a certain dictionary, whereas dynamic on-line dictionaries do not provide us with this information. If the Internet connection to that site is not reliable, then a query addressed to such a site might slow down the operation of *MoBiDictionary*. Only reliable sites with fast connection might be included.

4.2 Building user dictionaries

There is a strong wish among translators to build common terminological database on various fields of economy and science in Hungary: mostly English-Hungarian and German-Hungarian dictionaries. On a *MoBiDic* server anybody can create and build user dictionaries on any language that use any of the Latin, Cyrillic or Greek alphabets (full Unicode based representation is in preparation). A free and general on-line service is planned where the author's name and the address are parts of the newly created entries.

4.3 Concordance and parallel corpus module

It is also an old wish of translators to publish their translations for further processing. It would, however, arise authority problems, since the copyright owner of the translation is the translator, but he or she is not the copyright holder of the original text. Moreover, authors without any interest rarely agree on publishing their personal or internal official documents. There is, however, still a large amount of data that could be published, if there was an easy-to-handle (and free) site to store them and make possible to query them. *MoBiDictionary* is able to work with corpora, as well, so it might be a candidate for the above services.

4.4 Spell-checking and alternative suggestions for erroneous input

Querying of erroneous input results give usually no matches. In this case, two things used to happen. Either the user becomes disappointed knowing that he cannot supply the right spelling, or he/she accuses the dictionary because of missing important words. Both cases can be avoided supplying on line spell checking on input words. MorphoLogic has already developed spelling checkers for numerous languages, so it should be combined with the dictionary interface, as well.

5. Summary

MoBiDictionary, a new web site supporting translation activities has been introduced. It supports multi-dictionary access through linguistic stemming. One of the dictionaries called *MoBiGloss* has a special role: it is not a dictionarz but a connection to an (in principle) unlimited number of web based dictionaries without visiting them directly. The indirect connection means that *MoBiGloss* keeps in mind whether the user's actual query gives hits in certain dictionaries on the web, and it connects the user to the URL if and only if it contains hits to his/her query. Content of presently available Internet glossaries has also been discussed. In the near future, further extensions to the current services will be added.

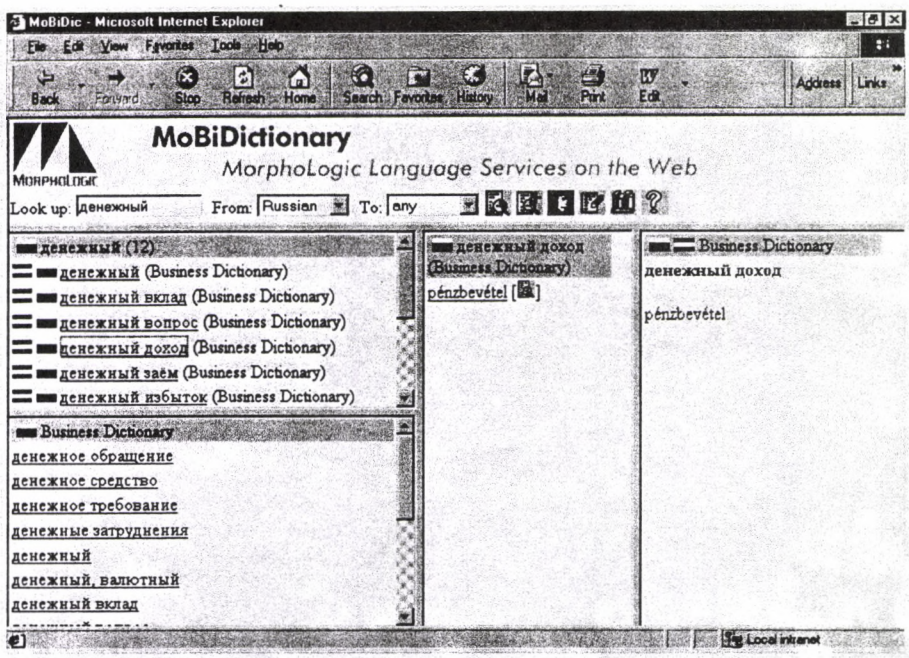


Figure 3. MoBiDictionary with Cyrillic text

6. References

Nerbonne, L. Karttunen, E. Paskaleva, G. Prószéky and T. Roosmaa (1997), Reading More into Foreign Languages. *Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington, 135–139.

Prószéky, G., Pál, M., Tihanyi, L. (1994) Humor-based Applications. *Proceedings of 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, 1241–1244

Prószéky, G. (1998), An Intelligent Multi-Dictionary Environment. *Proceedings of 17th International Conference on Computational Linguistics (COLING 98),* Montreal, 1067–1071.

The LVBAC Project: Contrastive Linguistics in a Bilingual Lexicon

J. TIÓ – J. M. COTS – M. SABATÉ – G. VÁZQUEZ –
F. MANYÀ – T. ALSINET

Abstract

This article is a presentation of the project Lexicó Verbal Bàsic Anglès--Català (LVBAC). The project is based on the development of a bilingual lexicon for computational purposes and machine translation. The project also deals with the essential problem of both syntactic and semantic desambiguation with the proposal of several strategies aimed at tackling this obstacle. As far as semantics is concerned, we suggest an analysis of semantic restrictions of the verbal arguments based on traits and values. In the program of analysis, we have used Prolog and chart, unification and semantic network algorithms.

In general, sentences containing figurative meanings or idioms have been discarded because these case would be more akin to other levels of linguistic description. We understand though that the line between non-figurative and figurative meaning is very fine and it is often difficult to perceive these differences. When this was the case, we have allowed our common sense and subjective criteria to be our guide.

These examples which, in theory, cover all the range of meanings of these verbs from the basic vocabulary, have been translated into English in every possible way, that is, using all the verbs which may be considered as synonymous. For this purpose, we have used the *Collins Cobuild* dictionary, the *Dictionary of Phrasal Verbs* (Courtney 1983) and the *Diccionari Català-Anglès* from the Enciclopèdia Catalana. As a result of this, most Catalan sentences produced multiple English translations. It is then our belief that our research has been comprehensive.

These pairs of sentences (Catalan original and English translation) have been analysed and classified according to the syntactic structure of the source language, i.e. Catalan. Next, we have studied the semantic mechanisms of desambiguation of some pairs of sentences, if such procedure was necessary.

3. Contrastivity

3.1 Syntax

From the empirical data obtained from the Catalan verbs analysed, we can appreciate the diversity of syntactic structures, that is, verbal subclasses (pronominal and non-pronominal verbs) and argumental structures.[3] For example, the Catalan verb *acabar* can be found in the following structures:

(1) NP1 V NP2	En Joan acaba els estudis
	John is finishing his studies
(2) NP V PPloc	El carrer acaba al riu
	The street ends in the river
(3) NP Vref	El temps s'acaba
	Time is running out
(4) NP1 Vref NP2	El nen s'ha acabat la sopa
	The boy has finished his soup

At this point, we need to point out that we have established two traits for the PP and they refer to the preposition or type of preposition they are introduced by. Thus, *tip* refers to the type of preposition, i.e., place, time, direction, origin, destination, etc., and *form* refers to its form and it is implemented whenever there is no other preposition available.

In (2) there is a place preposition (*a*), which may alternate with the preposition *en*; bearing this in mind, the entry that corresponds to this argument is defined as a PP with a place preposition (the pair *tip:loc*). In (10), however, the preposition *a* is the only one possible and therefore the entry is defined as a PP with the preposition *a* (the pair *form:a*). The lexicon as it stands shows the possibility of creating in the future a dependency network for the types of prepositions according to some classification proposals made by Trujillo (1995) and Lindstromberg (1998).

Some verbs show one only syntactic structure and are always translated the same in English. Such is the case of *amanir-to prepare, agrair-to thank, bordar-to bark, brodar-to embroider*, etc. Some of them have been illustrated below (5-8):

[3] Syntactic contrastivity has been already analysed by the research group (Tió *et al.* 1999).

(5) NP1 V NP2	La mare va amanir l'enciam
	The mother prepared the lettuce
(6) NP1 V NP2 PPa	En Joan li agraeix la invitació
	John is thanking him for the invitation
(7) NP V	Els gossos borden
	Dogs bark
(8) NP1 V NP2	L'àvia ha brodat els mocadors
	The Grandma has embroidered the handkerchiefs

Other verbs show the same translation in English in spite of having different structures. This can be seen in examples (9-11), in the case of *ajudar-to help* and (12-14) *aprendre-to learn*:

(9) NP1 V NP2	El nen va ajudar el seu pare
	The boy helped his father
(10) NP1 V NP2 PPa	El nen va ajudar el seu germà a fer els deures
	The boy helped his brother to do the homework
(11) NP1 V PPa	El nen va ajudar a aconseguir-ho
	The boy helped to do it
(12) NP1 V NP2	En Joan va aprendre la lliçó
	John learnt the lesson
(13) NP V PPde	En Joan va aprendre de solfa
	John learnt musical notation
(14) NP1 V NP2 PPde	En Joan va aprendre aquesta cançó de la seva mare
	John learnt this song from his mother

Other cases show always a one-to-one relationship, where a word-by-word translation would be more than enough. These cases are rare, though.

In other cases, very infrequently, we could see that their equivalence in English depends only on their syntactic structure and they have different equivalences depending on their structure. The translation of the verb *afanyar* in these sentences:

(15) NP1 V NP2	En Joan afanya el seu pa
	John earns his bread
(16)NP Vref	En Joan sempre s'afanya
	John is always hurrying
(17) NP Vref PPa	En Joan s'afanya a estudiar
	John studies hard

shows that the syntactic structure influences its English translation, which is different in each instance. In example (17) we can see the peculiar case of a Catalan verb reproduced in English as an adverbial; this is also the case of Catalan periphrases, which are more akin to the category of aspect. We have therefore left this issue aside for the time being for future research.

In relation to aspect, we have also observed that some tenses allow some changes in the argumental structure of the Catalan verb, which is translated differently in English. The example below illustrates this point:

(18) NP1 V NP2	En Joan ha caçat molts conills
	John has killed many rabbits

(19) NP V En Joan caça
John hunts

The second Catalan sentence is ambiguous: it may refer to a specific event or a quality attributed to the subject. The English translation of (19) refers to the latter. These instances are slightly unusual, though. Usually, a verb shows the same equivalence in English regardless of its tense. Similarly, the aspectual change of a verb does not usually entail changes in its argumental structure. In this example, there are two unusual circumstances: on the one hand, the English translation changes depending on whether the Catalan verb is in the present tense (19) or in any other tense (18). On the other, NP2 has been elided. In order to solve those instances illustrated in (19), we have thought about incorporating a variable on tenses in the description of the lexical entry.

Just as changes occur whenever NP2 has been elided, as in the above example, there are other changes which involve loss of arguments and/or change of position (passive, anticausative, medium, etc.). These variations are regular and therefore lexical rules may be applied to types of verbs. Although this has not been studied yet, it seems possible that structural changes result in a different English translation. This occurrence would call for a separate lexical entry.

From the analysis of our corpus of examples, we have observed that the most frequent cases are one only syntactic structure in Catalan with many different equivalent translations in English. This case will be developed in the next section.

3.2 Semantics

Whenever we needed semantic concepts to obtain a translation in English, using the popular trait *isa*, the nucleus of the semantic networks, often sufficed. For example, in the following sentences which include the verb *abocar*:

(20) NP1 V NP2 En Joan aboca un cep
John is planting a vine
(21) NP1 V NP2 En Joan aboca el cistell
John is emptying out the basket
John is tipping up the basket

we can see that applying *isa:vegetal* to the argument NP2 in (20) and *isa:recipient* to the argument NP2 in (21), is enough to obtain the corresponding English equivalences.

In other instances, additional information on the traits was needed since *isa* was not sufficient or not good enough. For example, the first structure of the verb *acabar*, cited above -example (1)-, NP1 V NP2, shows the following information:

(22) NP1 V NP2 En Joan acaba els estudis
John is finishing his studies
John's studies are coming to an end
(23) En Joan acaba la lectura d'un llibre
John is finishing the reading of a book
(24) En Joan acaba el vestit
John is finishing the dress
(25) En Joan ha acabat el vinagre
John has run out of vinegar
John has used up the vinegar
(26) En Joan ha acabat els diners
John has run out of money

The desambiguation of knowledge which a Natural Language Processing program may need to determine when to translate *acabar* into *to finish* (22-24) rather than *to run out* (25-26) will have to include the corresponding selectional restrictions. Using the trait *isa* is not sufficient if we intend to desambiguate, because NP1 always has the value of *human*, whereas NP2 has, both in the case of *to finish*, for example, *el vestit* in (24), and in the case of *to run out*, for example, *els diners* in (26), the value of *object*.

In view of this, we need to add a trait which enables us to differentiate NP2. We have suggested the trait *qua* (as in quality). Therefore, the proper translation will be *to finish* when the NP2 has the quality of *unexhaustible* (*studies, reading a book, a dress*, etc.), whereas it will be translated as *to run out* when it has the quality of *exhaustible* (*vinegar, money*, etc.); on the other hand, whenever something is *exhaustible* because it is *consumable*, then its translation may also be *to use up*, as in the case of *vinegar* in (25); finally, whenever something *inexhaustible* is an *activity* (*isa*), as in the case of *studies*, then it can also be translated using the verbal construction *to come* (*to an end*), as seen in (22).

From the data we have collected so far, we cannot positively say that this characteristic (*qua*) may be organized into a network of dependencies, unlike the case of the trait *isa*. We have considered this possibility in the case of *consumable*, because it has inherited the trait of *exhaustible*.

Now, if we read the following examples:

(27) NP1 V NP2	El fuster ha aprimat els taulons
	The carpenter has planed the planks
(28) NP1 V NP2	El fuster ha aprimat els bastons
	The carpenter has whittled the sticks

The syntactic structure is the same in both instances. Also, the only difference for both of them is in their NP2, although both *plank* and *stick* share the same trait *isa*, that is the value *object* among others such as *matter* which would have the value of *wooden*, etc. It seems that the only thing that makes *aprimar* translate as *to plane* a (27) or *to whittle* a (28) is a concept related to the roundness or flatness of the object in question: if NP2 is analysed *fig:pla* (flat), then the English equivalent is *to plane*; if NP2 is analysed *fig:rodo* (round), then the verb will be translated as *to whittle*.

Other examples of this phenomenon are:

(29) NP V PPa	Els ulls li brillen
	His eyes are sparkling
	His eyes are shining
(30) NP V PPa	La cara li brilla
	His face is glowing

Like in the previous instances, these syntactic structures are identical and their differerences revolve around a very similar vocabulary. In this case, neither the trait *isa* nor the trait *pof* are of any help. Nor are any more helpful the traits we have introduced above: *qua* (*exhaustible/unexhaustible*) and *fig* (*flat/round*) are not applicable to these instances. In this case, the difference between them may be found in the type of substance, matt or shiny, of the NP: the *face* (29) would be a *matt substance* (*subs:mat*) and the *eyes* (30) would be a *shiny substance* (*subs:brillant*).

4. The analysis program

The analysis program has been done in Prolog. From the linguistic point of view, this program is based on the grammar of traits and unification. From the computational point of view, it is based on chart algorithms.

The program components are currently the Catalan and English lexica (the part that has been implemented is made up of the first 80 basic Catalan verbs), the morphological component of Catalan (based on Tió & Manyà (1993)), the syntactic component of Catalan (new) and the chart and unification algorithms and semantic networks.

The corpus of Catalan examples corresponding to the first 80 verbs implemented amounts to 700 and the program works satisfactorily in 95% of the examples. Translation of the verbal items into English shows a much lower percentage of correct hits because the semantic components of some nouns included in the lexicon still need to be defined.

References

- Assessoria de didàctica del català (1975). *Vocabulari bàsic infantil i d'adults*. Barcelona: Biblograf.

- Courtney, R. (1983). *Dictionary of Phrasal Verbs*. Harlow: Longman.

- Gazdar, G. & Mellish, C. (1989). *Natural Language Processing in Prolog. An Introduction to Computational Linguistics*. Wokingham, England: Addison-Wesley.

- Klavans, J. L. (1994). Visions of the Digital Library: Views on Using Computational Linguistics and Semantic Nets in Information Retrieval. In Zampolli, N. Calzolari and M. Palmer, eds., *Current Issues in Computational Linguistics: In Honour of Don Walker*. Pisa: Giardini & Kluwer, p. 227-236.

- Lenders, W. (1989). *Computergestutzte Verfahren zur semantischen Beschreibung von Sprache*. Handbuch Computerlinguistik. Berlin: De Gruyter, p. 231-244.

- Lindstromberg, S. (1998). *English Prepositions Explained*. Amsterdam/Philadelphia: John Benjamins.

- Tió, J. and Manyà, F. (1993). Ortografia catalana i lingüística computacional. *Sintagma* 5, p. 59-70.

- Tió, J., Sabaté, M. and Vázquez, G. (1999). Syntactic Mismatches between English and Catalan. *Perspectives: Studies in Translatology* (in press).

- Trujillo, A. (1995). Towards a Cross-Linguistically Valid Classification of Spatial Prepositions. *Machine Translation* 10, p. 93-141.

Automatic Diacritics Insertion in Romanian Texts

DAN TUFIŞ – ADRIAN CHIŢU

ABSTRACT

The problem of automatic insertion of diacritics into an electronic text is well justified for several languages. Even if the diacritical characters concerned are present in the extended 8-bit ASCII charset (as the case is with French) any 7-bit filtering transmission of such a text will corrupt it. The situation is even worse when the diacritic characters are not in the 8-bit ASCII charset. To find a way to automatizing the diacritics insertion is worthy not only for old valuable texts stored in electronic form, but also for contemporary electronic texts as they continue to be produced in non-diacritical form. The reasons for this could be many, including the lack of localised and standardised keyboards. Ergonomic factors can also be mentioned (if someone is supposed to press more than two keys to get a diacritical character, then, mainly in informal communication (e.g. e-mail), he/she will probably take the easiest one-stroke solution).

In Romanian, every second word might contain at least one diacritical character and for large texts that lack diacritics, to insert them manually is highly time-consuming, boring and error-prone. Unfortunately, the automatic recovery of diacritics is non deterministic whatever the language which uses them. Various approaches can be envisaged to solve this problem but, when speed is of the utmost importance, then the approaches to be made are few.

We present such an approach for Romanian language based on our recent results in probabilistic tagging technology.

Introduction

The problem of automatic insertion of diacritics into an electronic text is well justified for several languages. Even if the diacritical characters concerned are present in the extended 8-bit ASCII charset (as the case is with French) any 7-bit filtering transmission of such a text will corrupt it . The situation is even worse when the diacritic characters are not in the 8-bit ASCII charset.

Simard (1998) provides examples of how a French text could become hardly readable when subject to such a filtering (not uncommon to many older programs that are not "8-bit clean"). For instance, a transformation that would systematically strip the 8-th bit in an ISO-Latin French text, would transform "é" into "i", "è" into "h" and so on. Other programs will simply delete accented characters (probably the most acceptable distortion) or replace them by a unique character such as "?" (certainly the least acceptable distortion).

To find a way to automatizing the diacritics insertion is of interest not only for old valuable texts stored in electronic form, but also for contemporary electronic texts as they continue to be produced in non-diacritical form. The reasons for this could be many, including the lack of localised and standardised keyboards. Ergonomic factors can also be mentioned (if someone is supposed to press more than two keys to get a diacritical character, then, mainly in informal communication (e.g. e-mail), he/she will probably take the easiest one-stroke solution).

Unfortunately, the automatic recovery of diacritics is non deterministic whatever the language which uses them. Various approaches can be envisaged to solve this problem but, when speed is of the utmost importance, then the approaches to be made are few. We will present such an approach for Romanian language, based on recent advances in probabilistic tagging technology. Similar approaches have been proposed in (Simard, 1998) for French, in (El-Bèze et al 1994) (cf. Simard, 1998) also for French. Yarowsky (1994) addresses this problem for Spanish (mainly) and French but instead of POS tagging, he uses a decision-list framework which offers very satisfactory performance (speed & accuracy) in spite of a language model that "was admitedly quite weak: in the absence of a hand-tagged training corpus, he based his model on an *ad hoc* set of tags" (Simard, 1998).

2. Diacritics in Romanian- a few statistical data

Romanian language has 5 diacritical characters: *ă,â,î,ş* and *ţ* (plus their uppercase variants). A text missing the diacritics will make that these characters be usually substituted by the ASCII characters *a* (for both *ă* and *â*), *i*, *s* and *t* respectively. This happens, for instance, when exporting from a diacritics-aware text editor into a text format. For a significant part of the words initially containing diacritics, their recovering is deterministic, because the non-diacritical variants of those words are not legal lexems of Romanian. But in most of the cases, the absence of diacritics creates genuine ambiguity, hard to resolve sometimes even for a human (when given only a limited context).

Here are some examples of strings that if missing diacritics are not legal words of Romanian (the real word and its translation are specified between parantheses):

A) padure (*pădure*- a forest), tufis (*tufiş*- a bush), bat (*băţ*- a stick), cantar (*cântar*- a balance), carare (*cărare*- a pathway), macar (*măcar* – at least), fara (*fără* – without), etc.

We call such strings unambiguous stripped words, or *U-words*.

To exemplify the ambiguity caused by the lack of diacritics, let us consider the string *fata*. In a text where the diacritics were removed from, this string could stand for any of the following words:

B) *fata* – the girl, *fată* – a girl; or (about animals) gives birth , *fâța* – the quick-swimming little fish/the coquette, *fâță* – a quick-swimming little fish/a coquette, *fața* – the face, *față* – a face. *făta* – (about animals) to give birth; gave birth, *fătă* – (about animals) just gave birth.

All the strings of the *fata* type above (i.e which could stand for more than one diacritical or non-diacritical words) are referred to in the following as ambiguous stripped words, or *A-words*. The strings that are neither U-words nor A-words are simply referred to as words.

When analysing a text from where the diacritics are absent (we call it a *stripped text*), if one has the right dictionary (this subject will be addressed in the next section) about 75-78% of the normal text (that is the text containing diacritics) can be deterministically reconstructed: 55-58% of the strings are actual words (so, no diacritics are required), and about 20% are U-words. The remaining strings (A-words, unknown words and proper names, approx. 25%) require special processing. In Romanian, in the vast majority of cases, it will do, knowing the morpho-syntactic properties of an A-word as prescribed by its contextual occurrence, to pick up the right diacritical or non-diacritical variant of it. Therefore, using the probabilistic tagging technology is a natural option when dealing. observing response-time constraints, with large quantities of stripped texts.

According to (Simard, 1998), approximately 85% of the words in French arbitrary texts carry no accents at all and more than 50% of the rest can be deduced deterministically on the basis of unaccented form. Consequently, for French, "with the use of a good dictionary, accents can be restored to an unaccented text with a success rate of nearly 95%"(Simard, 1998).

As compared to French, Romanian makes more intensive use of diacritical signs and their absence creates much more difficulties.

The table in Figure 1 displays data we extracted from our register-diversified training corpora (Tufiş, 1999). The three registers considered here are: fiction, philosophy, and journalism.

Text type	Fiction	Philosophy	Journalism
Total number of tokens	118,357	135,341	92,667
Number of relevant tokens	101,706	114,515	77,446
Number of words (diacritics free)	56,519 (55.57%)	65,615 (57.30%)	43,183 (55.76%)
Number of U-words	20,810 (20.46%)	23,176 (20.24%)	13,605 (17.57%)
Number of A-words	22,845 (22.46%)	24,951 (21,79%)	16,836 (21.74%)
Number of unprocessed tokens	1,532 (1.50%)	773 (0.67%)	3,822 (4.93%)
Proper names:	1,481 (1.45%)	706 (0.62%)	3,177 (4.10%)
Unknown word:	51 (0.05%)	67 (0.05%)	645 (0.83%)

Figure 1: Words, U-words and A-words as counted in our corpora

As shown in the table above, in computing various percentage figures we excluded from the total number of tokens those which are insensitive to the presence or absence of diacritics: punctuation characters, numbers, and dates. The number of such tokens, as shown in Figure 1, counts between 14% and 16.4% of the total number of tokens as identified by the system tokenizer (see next sections). The last line in the Table shown in Figure 1 reveals that proper names and unknown words are not currently addressed. We will come back to this issue in the section on evaluation.

3. The general architecture of the automatic diacritics insertion program (DIAC)

As already suggested in the previous section, the basic components of the automatic diacritics insertion program (DIAC) are a tagger and a special dictionary.

The tagger QTAG* is a slightly adapted version of Oliver Mason's QTAG trigram tagger (http://www.clg.bham.ac.uk/tagger.html), modified so that to comply with our methodology called tiered tagging (and more recently tiered tagging with combined classifiers) (Tufiş & Mason, 1998. Tufiş 1998, Tufiş1999a, Tufiş1999b).

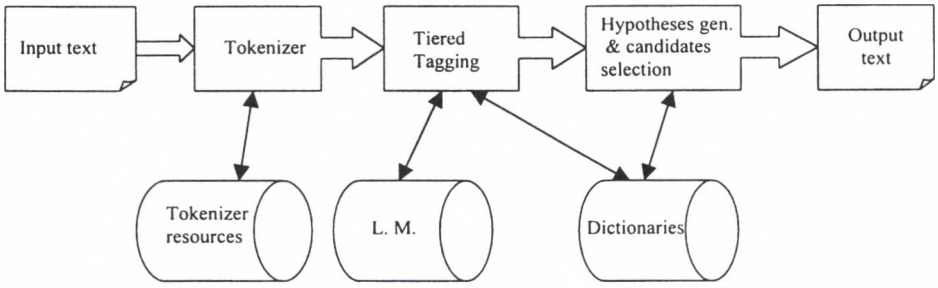


Figure 2: General Architecture of DIAC

Graphically represented in Figure 2, the processing flow is sketched below and detailed in the next sections:

a) a stripped input text is provided to the system, from either the keyboard or a file;
b) the input text is segmented into more lexical tokens according to the rules specified as external resources; a tokenized text is conveniently viewed as a vertical text, one token per line;
c) the tokenized text is subject to the tiered tagging process so that each token should be assigned a morpho-lexical description code; the tiered tagging process uses a language model (LM, built in a previous training step) and various dictionaries stored as external resources;
d) each token in the tagged text (the output of the tiered tagging) is associated with all possible diacritical and/or non-diacritical tokens (hypotheses generation) and based on the morpho-lexical tag, one form out of the candidate list is selected; when the information provided by the tag is not sufficient to make a decision, either lexical probabilities or some probabilistic preferences are used to make the final choice.

3.1 The tokenizer

The tokenizer is a program that identifies within the input text the elementary processing units called lexical tokens. A lexical token usually corresponds to the generally accepted idea of a word, namely a sequence of characters delimited by white spaces. However, several words may form a natural single unit (such as "*pentru că*" – because) or on the contrary, a sequence of characters delimited by white spaces may be split into distinct lexical units (such as "*dă-mi-le*" – you_(singular)_give to_me them = give them to me). The tokenizer also recognises, as one token, dates expressed in a large variety of formats (1 ianuarie, 1999; 01/01/99; 01-ian-99, etc), abbreviations (*dl, dna, dra, dr.* etc.), various types of punctuation, etc. Initially we used the MULTEXT tokenizer (developed by Philippe Di Cristo at LPL Aix-en-Provence)- a language independent and configurable tool. However, the price the MULTEXT tokenizer pays for its language independence and flexibility was considered too high. We have developed our own tokenizer which if not equally flexible, although still language independent, is at least 1000 times faster.

3.2 Tiered tagging

In highly inflectional languages, to encode the main morpho-lexical properties of the wordforms is to use a large system of description codes. The Multext European project in cooperation with EAGLES Lexical Specification Group developed a set of recommendations (Monachini & Calzolari, 1996) for the languages in Western Europe. Starting with these specifications, the Multext-East Copernicus project further developed them so that to account for the specificity of six other languages from Central and Eastern Europe – Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene – (Erjavec & Monachini 1997) and developed large compliant lexical resources (Tufiş et al. 1998). The set of morpho-syntactic descriptors (MSDs) specific to Romanian contains 618 codes. According to current practice, and in accordance with statistical rules, this number is too large. The larger the tagset, the larger the training corpora needed (Berger & al 1996). On the other hand, assuming that available training corpora are enough, LMs based on large tagsets would need large memory resources and given current hardware limitations, the overhead involved (required for memory management) may decrease the performance as low as to reach an unacceptable level. For instance, with about 700 tags, a LM of 300Mb or more would not be surprising (actually we got such a large transition table while training QTAG, see Tufiş & Mason (1998)) and on standard computers keeping it in RAM would be out of question. Apparently there are two solutions to get out of this deadlock: either to reduce the tagset to a manageable size and lose information or to modify the tagger with some extra-code taking care of data swapping and accept an unescapable (serious) response time degradation.

3.2.1 The basic methodology

The tiered tagging approach (Tufiş 1998, Tufiş & Mason, 1998) is a nice answer to this dilemma, solving the contradiction between a small tagset (needed for tagging accuracy with reasonably large training corpora, fast response time and limited computational resources) and a large tagset (needed for a finer-grained classification, not necessarily distributionally relevant but needed for the diacritics insertion).

With a small price in tagging accuracy (as compared to a reduced tagset approach), and practically no price in computational resources, it is possible to tag a text in terms of a large tagset by using LMs built for reduced tagsets and consequently requiring much smaller training corpora.

In general terms, tiered tagging uses a hidden reduced tagset (we call it C-tagset and in our case it contains 99 tags, including punctuation tags) based on which a LM is built. This LM serves for the first level of tagging. Then a post-processor deterministically replaces the tags from the small tagset by one or more (in our experiments rarely more than 2) tags from the large tagset (we call it MSD-tagset). The words becoming ambiguous after this replacement are more often than not the difficult cases in statistical disambiguation (such as determiners vs. possessive pronouns). They represent a small percentage (in our experiment, less than 10%) and are further processed by means of some very simple regular-expression rules. These rules investigate, depending on the ambiguity class, left, right or both contexts within a limited distance (in our experiment never exceeding 4 words in one direction) for a disambiguating tag or word-form. The success rate of this second phase was higher than 98%, that means that the additional error introduced by the C-tag to MSD-tagset process is less than 0.2% . Certainly, the reduced and the extended tagsets have to be in a specific relation (C-tagset should subsume MSD-tagset). As general performance of tiered tagging is practically the result of the hidden layer tagging (actually the proper statistical disambiguation), the design of the C-tagset is of utmost importance and therefore it has been very carefully designed to observe the limits of probabilistic tagging (for instance, we dropped all the attributes that could not be distinguished based on distributional criteria or most of the attributes/values which are implied by

the values of other attributes and the wordform itself). Another design criterion for the C-tagset was what we called *maximum-recoverability* property, defined as follows:

Given:a word W in a tagged text (C-tagset). the MSD-ambiguity class of W (encoded in the lexicon), the tag T from C-tagset which a tagger assigned to W, and the list of MSD-tags that are mapped onto T, then the mapping from T to the correct MSD tag should be *almost deterministic*. In our case, the *almost deterministic* requirement translated into a mapping that is unique in almost 90% instances. For the remaining ambiguous cases there will never be more than 2 MSDs to consider.

As a direct consequence of the efforts invested in the C-tagset design, the average accuracy of QTAG* trigram tagger on various test texts (no one less than 20,000 words, unseen before and from various registers) is constantly well beyond 98%. By building a combined classifier out of the basic 3 language models (learnt from the three training corpora mentioned before) we managed to constantly increase the accuracy above 98.5% (Tufiş, 1999b).

According to our intensive empirical evaluations and experiments with numerous C-tagsets and various taggers, we dare say that a simple tagger working with an appropriately designed C-tagset will more often than not outscore a sophisticated tagger working with a badly designed set of tags. Once a C-tagset properly designed (i.e. both correct from a distributional point of view and maximally recoverable), a tiered-tagging approach to the large tagsets problem is simpler, cheaper, faster and more accurate than that of developing very large training corpora, intensive training and using sophisticated smoothing techniques.

3.2.2. Adapting the training resources and the dictionary for the DIAC system

For the purpose of the DIAC system various modifications imposed on our basic tagging resources. First, the tagger's dictionary, containing words associated with various tags and their lexical probabilities[1] in the following format: <$word_k$ c-tag_1:p_1 c-tag_2:p_2 ... c-tag_n:p_n> had to be diacritics stripped-off. As discussed in section 2, the removal of diacritics can create either U-words or A-words (if the word contains no diacritics, nothing happens on its associated entry). For those entries where if removing the diacritics a U-word will result, the initial word is replaced by the corresponding U-word but the rest of the entry (i.e. c-tag_1:p_1 c-tag_2:p_2 ... c-tag_n:p_n) remains the same. For the entries of which keywords are mapped onto a A-word, the resulting entry is obtained by a merging procedure.

The training corpora (vertical texts with two columns: <word><c-tag>) are created from the original training corpora by simply dropping off the diacritical characters. From these diacritics-less training corpora, the learning part of QTAG tagger creates the language models (LM) to be used in tagging new stripped texts.

Beside the tagger's dictionary, DIAC makes use of a mapping dictionary (implemented as a hash table), the structure of which is described by the BNF grammar below:

<entry>::= <non-diacritical entry>| <U-entry>|<A-entry>
<non-diacritical entry>::=<word> 0 <MSD>
<U-entry>::=<U-word> 1 <real_word> <MSD>
<A-entry>::= <A-word> {<real_word> <MSD>}$^+$

One should notice that the tagger's dictionary contains tags in the reduced tagset (C-tagset) while the mapping dictionary contains the morpho-syntactic descriptors (MSDs). The tagset mapping

[1] Actually they are not exactly lexical probabilities but occurrence counts extracted from the training corpora.

table required by the tiered-tagging methodology, which associates for each tag in C-tagset the set of MSDs it subsumes, is not affected by the specific requirements of the DIAC system.

3.3 The hypotheses generator and the candidate selection

The entry to this module is a sequence of pairs <*token c-tag*> where the *c-tag* belongs to the reduced C-tagset and *token* is a U-word, an A-word or a simple word (that is neither U-word nor A-word, see section 2). The *token* is subject to further processing provided it is not a number, a date, an abbreviation or a proper name. In such a case, it is left untouched and for proper names (which in Romanian frequently contains diacritics) a warning message is written in the log file.

The hypotheses generator is a simple look-up program that searches the mapping dictionary for the entry associated with *token* (used as a key in the hash table). If an entry is not found, again a warning message is written in the log file. Otherwise, the returned entry contains the real word candidates out of which one would replace (if the case) the current *token*:

- If the entry corresponds to a non-diacritical entry (<word> 0 <MSD>), then the current token remains unchanged.
- If the returned entry corresponds to a U-entry (<U-word> 1 <real_word> <MSD>), the current token (U-word) is replaced by its unique diacritical form (<real_word>).
- When the returned entry corresponds to a A-entry (<A-word> {<real_word> <MSD>}$^+$), the candidate selection is done according to the next steps:
    1. the *tag* associated with *token* is expanded to the appropriate MSD (this is part of the second phase in tiered-tagging (Tufiş & Mason 1998).
    2. If the resulted MSD appears only once in the A-entry, then the <real_word> associated with this MSD is the selected candidate.
    3. The other case, when the MSD appears more than once in the A-entry, it is solved empirically, based on user preferences (if they are defined) or occurrence frequency of the competing <real_word>s (this implies several look-ups in the tagger dictionary). At present, it is possible to express user preferences only with respect to the tense of the main verbs (present or past). This step is responsible for almost 40% of the errors in automatic insertion of diacritics.

4. Evaluation

The DIAC program was evaluated first in an idealised setting, namely assuming a perfect tagging (that is 100% accurate). We used the training corpora, stripped of the diacritics and gave the resulted texts as input to the hypotheses generator and the candidate selection module. We also played with various tense preferences until we reached a maximum accuracy for each type of text. The results are shown in Figure 3.

Corpus	Fiction	Philosophy	Journalism
Number of tokens	118,357	135,341	92,667
Number of relevant tokens	101,706	114,515	77,446
Number of errors	631	423	1177
Idealised Setting Accuracy	99.38	99.63%	98.48%

Figure 3: Estimation of the accuracy of the hypotheses generator and the candidate selection module (assuming no tagging errors)

The high number of errors in the "Journalism" corpus is explained by a large number of Romanian proper names (3,177) out of which 718 contained diacritics. Because, as specified before, the current version does not process the proper names, all the 718 proper names counted as errors. On the other hand, the Fiction corpus (Orwell's "1984") and the Philosophy corpus (Plato's "The Republic") contained no occurrence of a proper name using diacritics.

The real setting evaluation consisted in applying the entire processing chain (that is including tagging). We used 4 language models (the three basic ones extracted from the training corpora and a combined classifier) observing the restriction that a given LM should not be applied to the text from which it was learnt. The errors introduced by the tagging process decreased the accuracy as compared with the idealised setting by an average of 1.3%, but it is worth mentioning that the final error rate was significantly smaller that the intermediary error rate of the tagging phase. The explanation for this is given by the fact that some errors of the tagger are not important for the diacritics insertion process. For instance, if a token is a definite feminine noun but is tagged as a definite feminine adjective this is certainly a tagging error. However, for the diacritics insertion what matters is only the definiteness value irrespective of the noun or adjective distinction. A similar case (quite frequent) is the mistagging of verbs of the first conjugation (that is ending in "a" in the infinitive) main_verb/indicative/present/3rd-person/ singular as main_verb/indicative/ simple_perfect/3rd-person/singular (or viceversa). In speech these two forms are distinguished by a different accent position, but in written form they are graphically indiscernible (in written Romanian the accent is not usually indicated): cântă (pronounced as c`ântă ) -s/he sings versus cântă (pronounced as cânt`ă) -s/he has sung.

The table in Figure 4 shows the intermediary tagger accuracy and the final DIAC accuracy for various language models and texts combinations.

LM / Test corpus	Fiction		Philosophy		Journalism		Combined Classifier	
	Tagger accuracy	DIAC accuracy	Tagger accuracy	DIAC accuracy	Tagger accuracy	DIAC accuracy	Tagger accuracy	DIAC accuracy
Fiction			96.42%	97.53%	96.33%	97.55%	97.28	97.73%
Philosophy	95.82%	98.27%			96.05%	98.22%	97.05%	98.47%
Journalism	96.35%	97.12%	96.61%	97.08%			97.44%	97.36%

Figure 4: Intermediary (tagging) accuracy and final (DIAC) accuracy

There is one very important observation to be made, concerning the tagging accuracy. In previous papers on tiered tagging and combined language classifier methodology (Tufiş 1999a, b) we mentioned much better performances for the disambiguation process (well beyond 98,5%) which are apparently contradicted by the results shown in Figure 4. The reason for the significant observed difference is fully justified by spurious ambiguity, hard to resolve, created by the diacritics removal. For instance, the word "ca" is either adverb (Rgp) or conjuction (Csssp) while the word "că" is just a conjunction (Csssp). Removing the diacritics, it will result an A-word "ca" with the ambiguity class (Rgp Csssp) which is notoriouly difficult to solve. In (Tufiş & Mason, 1998) we discussed this ambiguity for the words "şi" and "ca" and showed that in order to avoid this very annoying source of error we introduced a "port-manteau" tag (RC) just for these two words. We argued there, that the distinction between the adverb and conjunction interpretation for the two words is so fine that most native (non-linguists) Romanians would not differentiate them. Also, we gave evidence that the distributions of the two interpretations for the two words are practically the same. In the context of DIAC application, however, this distinction has to be made for distinguishing diacritical forms "ca" and "că" as well as other compounds based on the two words ("pentru ca" and "pentru că"). Just this case is responsible for almost 1% of the non-diacritical tagger! Another supplementary error source generated by the diacritics absence is related to the vanished graphical difference between indefinite feminine nouns/adjectives, singular, direct case (which always ends in "ă") and definite feminine nouns/adjectives, singular, direct case (which always ends in "a"). In many cases

this distinction may still be recovered, because a definite noun/adjective cannot co-occur with an indefinite article (always proclitic). However, the absence of the indefinite article before a non-diacritical form of a feminine noun/adjectives, singular, direct case says nothing about its definiteness. Due to this case, the non-diacritical tagger makes about 0.3% more errors. Finally, some distinctions we ignored in the verbal tags for regular texts (containing diacritics) cannot be ignored anymore, because they impose whether a diacritic is inserted or not. As an example we can mention the finer distinction between imperfect and simple perfect needed to turn the A-word "pleca" into "plecă" (s/he left) or "pleca" (s/he was leaving). On the other hand, it is worth mentioning that several distinctions made in the regular tagset appear to be too fine-grained for tagging stripped Romanian texts for the purpose of automatic insertion of diacritics. This fact suggests that experimenting with various tagsets, specialised for the application considered here, might be rewarding.

One should notice that the accuracy of the combined classifier tagging (see Figure 4) is significantly better that of any individual language model and the improvement in the DIAC environment is more consistent than when tagging diacritical texts. This happens because the disagreement ratio between the tags assigned by individual classifiers (LMs) increases when the diacritics are stripped-off, but the correct assignment continues to be among the proposals out of the combined classifier has to make the final decision.

5. Implementation, current limitations and further work

The DIAC system is implemented under Unix in a client-server architecture (but a stand-alone version is easily simulated via localhost). The server is actually written in C and Perl, but soon it will be rewritten in Java. The client is written in Java. There are various selectable options of representing the automatically inserted diacritics: ISOLatin2, SGML entities, tex-like, etc.). In the current version some tokens after tagging are further ignored. They are as follows:

a) non-word tokens (numbers, punctuations, dates, special characters)
b) words not in the dictionary; although (via the guesser) an unknown word could be assigned a correct tag, not having it in the dictionary prevents further decisions on whether and where diacritics should be inserted. The word remains unchanged, but a warning is recorded in a log file.
c) words which are tagged as proper names; given the proper names are not normally recorded into a dictionary, they are dealt with in a similar way to the unknown words. However this is temporary, since building a specialised proper names dictionary containing only A-words and U-words is an ongoing project. As shown before, many Romanian proper names contain diacritics and not treating them explains the significant error rare between newspaper texts and literature translated from English.

The mass data (tagger dictionary and the mapping dictionary) are encoded as hash tables, using UNIX gdbm. We plan migrating this information into a proper database with regular database management facilities. The envisioned DBMS is ORACLE8™. The DIAC system will be made publicly available on the web as a mail-service. Further experiments will investigate the effect of application-oriented customisation of the tagset on the accuracy of the automatic insertion of diacritics.

Acknowledgements

The work reported here built on the main results of the Multext-East European project (COP106/1995) and was partly funded by a grant of the Romanian Academy (GAR#188/1998)

References

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, J. H., Petkevič, V., and Tufiş, D. (1998). "Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages" *In Proceedings of COLING-ACL '98,* Montreal, Canada, 315-319

Berger, A., L., Della Pietra, S., A., Della Pietra, V., J. (1996): A Maximum Entropy Approach to Natural Language Processing in *Computational Linguistics*, vol. 22, no. 1 (pp. 39-72), March 1996

Bèze et all 1994

Monachini, M. & Calzolari, N. (Eds.) (1996) EAGLES Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora A Common Proposal and Applications to European Languages (EAG---CLWG---MORPHSYN/R August, 1996 (http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/morphsyn.html)

Erjavec, T. and Monachini, M. (Eds.) (1997). *Specifications and Notation for Lexicon Encoding.* Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/mte-d11f/.

Simard, M. (1998) "Automatic Insertion of Accents in French Texts." *In Ide & Vuotilainen (eds) Proceedings of the Third Conference on Empirical Methods in Natural Language Processing* . Granada, Spain, 27-35

Tufiş, D. (1998) "Tiered Tagging", Research Report no. 32, June, 1998, RACAI, Bucharest, 72pp (in Romanian)

Tufiş, D., Mason O. (1998). "Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger" *In Proceedings of First International Conference on Language Resources and Evaluation,* Granada, Spain, 589-596

Tufiş, D., Ide, N., Erjavec T. (1998). "Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages" In *Proceedings of First International Conference on Language Resources and Evaluation,* Granada, Spain, 233-240

Tufiş, D. (1999a) "Combined Language Models and Tiered Tagging for Highly Inflectional Languages", forthcoming

Tufiş, D. (1999b) "TT-CLAM: A Novel Approach in Statistical Morpho-Lexical Disambiguation" In *Proceedings of the 12th CSCS,* Bucharest, June 1999 (to appear)

Yarowsky (1994)

Looking in before Looking out: Internal Selection Criteria in a Corpus of Plant Biology

GEOFFREY C. WILLIAMS

Abstract.

Representativity has always been a problem in corpus linguistics and is all the more acute in specialised corpora which tend to be small. Domain-specific corpora are often referred to in terms of sublangages. This text claims that sublanguage is too vague a notion to be of use in building representative specialised corpora. The notion of Discourse Community which takes into account the rhetorical choices of a research community, and the subsequent lexico-grammatical choices used in communication, gives a clearer picture of the interaction between disciplines inherent in all research. It is claimed that specialised corpora can be built using strict external criteria in order to reflect the language of a Discourse Community and the terminological influences of related fields. However, corpora based on purely external criteria will contain a great deal of noise. To make these corpora exploitable, it is necessary to apply internal lexical based criteria in order to isolate themes within the corpus. Lexical criteria are preferred over grammatical ones as specialised corpora tend to represent one particular genre. It is shown how marking up certain terms using SGML/TEI tags can enable the delimitation of themes. The use of internal lexical criteria is demonstrated on a corpus of research articles in plant biology, more particularly the biology of parasitic plants viewed from the viewpoint of plant physiology and plant molecular biology.

Looking in before looking out: Internal selection criteria in a corpus of plant biology.

Introduction

Representativity is a central notion in corpus linguistics. This is often achieved through created ever bigger corpora, yet, size remains an insoluble question, large being the only answer (Sinclair 1996). In the generalist mega-corpora, such as the BNC, balance of content is justified on statistical grounds. Whilst it may be possible to claim a degree of representativity in large corpora, on the basis that the sum is greater than the parts, it would be difficult to claim that the scientific elements of the BNC are truly representative of scientific English. Taken out of the main corpus, the 'balance' no longer exists. In scientific corpora, 'representativity' is complex, all depends on what we claim to be representing when we select texts, an awareness of noise from material not central to the theme under study is paramount.

There is thus need for new criteria in the building of scientific corpora. The need for such corpora is paramount. (Meyer and Macktintosh 1996) demonstrate that although corpora are widely available in general lexicography, they are far and few between in terminography. The main constraints in building the necessary corpora are time and domain specificity. Corpus building is a very lengthy and problematic process, especially in terminography which must keep abreast of current developments in a field. Domain specificity is essential for a corpus to be truly representative of the field being described, but such specificity is often hampered by the constraints of availability of material. Hence, although much work has been done on tools for terminology extraction using collocation (i.e. Daille 1995) and term variation (Jacquemin 1996), for such tools to be tested, applied and perfected domain representative corpora are desperately needed.

Unlike with general language corpora, scientific and research corpora must be composed of the complete article, rather than a sample. The article consists of a number of easily definable subsections each with its particular rhetorical choices (Bhatia 1993). To choose one section, such as the abstract, simply because of easy availability, does not give representativeness. It is true that discourse can be segmented by machine for computer analysis of the subsections (Berber Sardinha 1997), but, for terminological extraction, the text as a whole is required so that the variants used in subsections may be isolated. Such a requirement precludes the building of corpora from abstracts only, appealing as it may be due to their copyright status. The choice should be made by selecting a optimal number of documents to illustrate a given domain. This is generally done using external criteria with reference to the notion of sublanguage.

This study is an attempt to isolate selective factors to permit the construction of fine-grain selective criteria within scientific corpora. This is done by a microscopic study of texts using collocation as a selective tool, followed by macroscopic analysis on a 400 000 word corpus of plant biology research articles.

Sublanguage and Discourse Community

Sublanguages, following the terminology of (Harris 1968), are essentially a generative invention designed to overcome the need to explain the restricted syntactic choices operative in certain domains. Whilst the need for special corpora in NLP has led to closer definition of the term (McNaught 1992), sublanguages remain difficult to tie down, the term often being little more than

shorthand for describing the usage of a particular domain with its particular lexico-grammatical choices. This study claims that such an isolationist approach cannot lead to representative scientific corpora. Clear frontiers between domain-specific language and so-called general language simply do not exist, what we have is a cline of restrictivity in which one 'language' merges into another in line with the rhetorical needs of the writer. A better approach to corpus design would be to take into account the purposes of communication. This can be accomplished by reference to the notion of discourse community (DC) (Swales 1990). A researcher may belong to several discourse communities, it is thus necessary to couple our external selection criteria, based on genre and domain, with internal criteria that illustrate not only the concepts under study, but the approach adopted to those concepts.

According to (Swales 1990: 26) a discourse community

> utilizes and hence possesses one or more genre in the communicative furtherance of its aims.

This means that rather than saying that a sublanguage exists and then looking for it, we begin from the viewpoint that the DC exists only for the purposes of its owners, the rest of us being but observers. The norms of that community are agreed tacitly, but not arbitrarily, being the results of the shared publication norms of the other members of the DC. Consequently selection criteria must try to reflect this belongingness. Taking one journal as the basis of a monosource corpus is obviously interesting both for the variety of language used and for ease of negotiation in terms of copyright. However, unless the community under question is the sole user of that organ of communication, and that it does not publish for purposes of the DC outside that organ, a single journal will inevitably represent a variety of DCs, both field-centred and topic-centred.

One corpus built on the DC principle and using a variety of sources is the PSC corpus investigating the language of cancer research (Gledhill 1994). Gledhill gives his criteria as being

1. Internal cohesion. - texts produced by those researchers who represent a theme.
2. External cohesion - texts to which the research team has been exposed.
3. Global coherence - texts and journals that treat a given theme but do not directly refer to the precise topic under study

(adapted from Gledhill 1997)

The advantage of such an approach is to locate a theme within the world of science to which it belongs, thereby demonstrating the interests of a discourse community and the influences on that community. The disadvantage, in terms of lexicography or terminology, is the considerable noise around the theme. We are at the limits of external selection criteria. To go further it will be necessary to isolate internal selection criteria, criteria that would allow the isolation of precise themes within the corpus. Biber's work on genre (Biber 1988) has been highly influential in breaking down content by genre, such an approach is not, however, applicable here as specialised corpora work within a single genre. At this level it seems unlikely that grammatical internal criteria would be applicable. On the basis that a DC has its own lexis (Swales op.cit 26), criteria would of necessity be lexical. In this case, work to find internal selection criteria is being carried out on a corpus of plant biology, the results therefore can only be applied in this domain, only the method is reusable.

The BIVEG Corpus in Plant Biology

The BIVEG corpus (Williams 1998 : 166-167) has been built on similar lines to that of Gledhill, in that the texts have been selected with the assistance of researchers. However, in this case three clear poles can be isolated, texts referring to molecular biology, texts from the field of plant physiology and a mixed selection from a conference on parasitic plants, the main theme of this corpus. Using this corpus, the aim is to find factors to allow selection of texts dealing with parasitic plants, then to isolate the texts emanating from the two main approaches being used, and to gradually refine the selection process before enlarging the corpus. Once factors have been isolated it should then be possible to obtain material in machine readable format from publishers, using internal selection criteria to select texts in line with individual themes.

Whilst the physiology and molecular subsets consist of texts drawn from a variety of journals and do not all deal with parasitic plants, those from the proceedings of the 1996 Conference on Parasitic Plants in Cordoba (Moreno et al 1996) could be seen as the state of the art in one topic-specific domain, and, therefore, representative of the DC at work. However, a conference proceedings can only reflect the papers approved by the selection committee, to accept this as the basis for the extraction of lexis would be to ignore the constituent themes. In this instance the papers were put into categories with expert assistance, the object here is to verify the objectivity of this categorisation, a task which can only be done through reference to a wider corpus.

The early criticisms of corpora by the generativist linguists still have some value. A corpus cannot contain the whole truth and nothing but the truth, it remains necessary to look beyond the immediate context. To do this, we need to build larger corpora, clearly distinguishing the subsets within than corpus. It may well be that the information we wish to include in a dictionary will be only relevant to the contexts of a particular domain, but in order to understand those contexts we must look beyond. An example from the corpus under study is the article by (Sigiura 1992) on the chloroplast genome. This text does not deal with parasitic plants as such, it is, however, the state of the art in chloroplast knowledge and therefore essential for an understanding of the development of the molecular approach to parasitic plant biology.

The first stage in corpus exploitation is to divide the texts into different themes. There is obviously a multitude of ways of doing this by discipline, centre of interest, even plant type. What was done here was to follow the advice of the researchers and divide the texts according to the two main disciplines under study, plant physiology and plant molecular biology. This was relatively easy, what was less so was categorising the texts in the Cordoba proceedings. This was done with reference to 5 themes, the two aforementioned, general plant biology, systematics and agronomy. Such subgroups are inevitably subjective, the particular interests of our experts implies a certain bias in classification. This is compounded by the fact that there are no rigid boundaries between so-called sublanguages, a text may be part of several DCs or themes depending on the intentions and background of the authors. In this study, the subjective nature of the groupings are irrelevant as once factors have been found and applied the boundaries can, and will, be redrawn on the basis of lexical content.

Main theme : Parasitic Plants.

The central DC that this corpus has been established to explore is that of parasitic plants, particularly root parasites. These are plants, classed as either wild plants or weeds, that live off a host plant, the best known to most people is probably mistletoe, *Viscum album*, common on apple and oak trees. Less known are the root parasites, as the genera *Striga, Cuscuta and Orobanche*, which can be

major agricultural pests on a variety of staple crops including maize, rice and tomato. One of the particularities of these plants is that they only germinate in the presence of a host, and once attached are difficult to destroy with herbicides, having no roots of their own.

In order to select texts on the main theme, the obvious approach would be to use the term 'parasitic plant' as a key word. However, this term only occurs 22 times (0.005%) in 15 texts making it a very poor selector. The word 'parasite' or even *parasit*, so as to catch 'holoparasite', 'parasitising' etc are no better as they are to be found in 83 texts with relative frequencies varying from 0.019 to 2.102 for an overall frequency in the corpus of 0.281. If using such 'key words' is too simplistic an approach, another way to tackle the problem must be to look at the plants themselves.

As the parasitic plants belong to a limited number of genera, with only a few common names in use, it is relatively easy task to mark them up. The corpus being encoded in conformity with the TEI Guidelines, the element <TERM> was selected using the attribute 'type' with a value of 'pp' for parasitic plant. The host plants were also marked up with the attribute value 'hp'. This is less easy as the hosts are a more open set. Once marked up all occurrences of the term type="pp" could be called up and those texts selected. This led to the selection of 87 of the 155 texts that form the corpus. The relative frequencies were calculated for each text and the scores represented as log10, as can be seen on graph 1.

The next stage was to isolate the collocates of 'type="pp" to see whether these could serve as further selection factors. Collocations were calculated using mutual information scores (Church et al. 1990, 1994) applied within the subset of 87 articles. Taking the collocates of these 87 texts it was then possible to calculate the relative frequency of the collocates which, combined with the type score, allowed the elimination of two borderline texts. The two texts eliminated were BV009GEN and BV015GEN. In these two texts, the plants were simply cited as examples in a very different context. As can be seen on the graph, the collocates make poor selectors as there is no significant differences between the texts. It is on the "pp" content that the two eliminated texts stand out. Mutual Information may well not be appropriate in this case as the grouping of the plant names unbalances the frequency relationships in the corpus. Nevertheless the collocates do give valuable information about the corpus as will be seen later.

So far nothing revolutionary has been done, except that 44% of the corpus has been eliminated. Following the same selection procedure with type="hp" (host plant) showed that 122 of the texts deal with the host plants, thereby revealing the existence of two DCs, one concerned with certain aspects of the host plants and another, partially overlapping, that deals with the relationship between host and parasite. It remains to discover the relationship between these two DCs, and the DCs represented by the other texts in the full corpus. The now 85 texts of the parasitic plant subset cannot be said to be representative of the entire parasitic plant DC, but only of part of it. Which part?

The molecular subset

As stated earlier, this corpus has been built following two domains, plant physiology and plant molecular biology, so the next task must be to isolate the factors predictive of these domains and then check the results against the expert's intuition. For molecular biology, the simple starting point is by reference to DNA in its different forms, RNA and lexis based around the concept 'gene'. For the moment the genes themselves have not been marked up as the isolating of the individual gene names is a long task requiring intense human intervention and is, as of present, unfinished.

The same approach for text extraction was used, both for the full corpus (data not shown) and for the pp subset (graph 2). What emerged was that certain texts classified as molecular in the full corpus were rejected, others classed under physiology had a strong molecular element. Within the Conference subset, all those designated as belonging to molecular biology were selected, others too, indicating that these had either been wrongly classified or that they partake in several themes.

Main Theme and Poles of Interest

As mentioned earlier, the use of collocates as defining factors did not prove conclusive. What did become apparent is that a number of essential sub-themes exist, themes that are of particular interest within a domain, whilst not rising to the level of DC. These sub-themes can be referred to as poles of interest or simply 'poles'. They are generally too wide a subject area to be considered a DC, but do represent essential areas of research that feed into the main topic.

Graph 3 shows the relative frequency of collocates for the pp subset. Some stand out as having a particularly high frequence, others not. Amongst the lower frequencies we find dodder, the common name for Orobanche, a parasite, and sunflower, a host plant. We also find haustorium, the nodule formed where a parasite fixes to the root of its host. The latter is an essential element in defining a parasitic plant, but not considered of outstanding interest by the authors contributing to this subset who take its presence for granted. Three poles do stand out, species, control, and germination. It is obvious that these do need to be explored to find out their significance to the corpus as a whole, representing as they do potential lead-ins to collocational networks (Williams 1998). Obviously certain poles concern one discipline more than another, control in particular concerns almost exclusively texts from the physiology and agronomy subsets, concerned with the effects of herbicides and limiting parasite attack. Species is altogether a more general concept concerning all the texts, but with a particular interest for comparative studies in systematics. Germination is a particularly strong pole, especially in physiology, a fact that is easily explainable given the germination requirements of parasitic plants. This inevitably brings in texts concerning this phenomenon from outside the subset. In other words looking in will isolate the phenomenon as of significance within a certain domain, looking out will allow us to understanding the terminology involved.

Restricting our subset to the molecular section revealed the importance of genomic influence, BV009GEN whilst being eliminated as marginal in terms of parasitic plant biology provides an important link here. BV015GEN, also considered as marginal has a particular interest in plastids, an area of particular interest within the molecular biology of parasitic plants, one that links out to the text of Sigiura mentioned earlier.

What these few examples demonstrate is that a corpus built on a purely statistical basis around a so-called sublanguage will inevitably miss much of interest to the terminologist. Using Gledhill's model, and thereby taking into account the influences that come to bear on a DC, will bring in this missing material, but also dilutes the central theme. The only way to find a compromise between these two extremes is to look at the lexical factors within the corpus to see exactly what is there. By making the bias explicit, we enter the domain of justification of content rather than abstract representativity.

Conclusion

By gradual refining of the selective items we are able to clearly delineate themes in the corpus. These items can then be used as selection factors in a larger, more representative, corpus, exploited as a

whole, and with reference to the interlocking thematic sub-corpora of which it is composed. The end result would be a more rigorous internal selection within the corpus.

This call for fine-grain scientific corpora does not dispute the validity of current extraction methods, what it does do is call for a clarification in the claims made by lexicographers as to the field and sub-fields covered in specialised dictionaries. To rely on corpora without taking a close look at what they represent may well lead to the domination of one field over another, an unwitting pulling of the sheets to one side of the terminological bed.

Representativity is a notoriously vague notion, recourse to statistical techniques no matter how sophisticated will not raise this vagueness. Representativity can only be spoken of in terms of representing something, something that is explicitly declared in line with selection criteria. Unfortunately tying down that something is far from easy, all the more so when our corpus-building activities are so hampered in terms of availability of data, particularly due to copyright restrictions. Maybe in building and, above all, exploiting corpora we should bear in mind the importance not just of representativity, but also of justification. All corpus building requires the making of choices, we should be clearer not only as to the nature of those choices, but also as to the contents of our corpus. This cannot be done purely by the use of external criteria, nor the use of transposable internal criteria, but requires looking closely at our material with human eyes so as to see what is within before attempting to look out.

References

Berber Sardinha, A. P. (1997). *Automatic Identification of Segments in Written Texts*. Unpublished PhD Thesis. English Department/AELSU, University of Liverpool. UK.

Bhatia. V.K. (1993). *Analysing Genre: Language use in professional settings*. Longman : Harlow

Biber. Douglas. (1988). *Variation across speech and writing*. Cambridge: CUP.

Church, K. and Hanks, P. (1990). *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics Vol 16/1 :22-29.

Church, K., Gale, W., Hanks, P., Hindle, D., Moon, R. 1994. Lexical Substitutability in Atkins and Zampolli (1994). *Computational Approaches to the Lexicon*. Clarendon Press: Oxford.153-177.

Daille, B. (1995).*Combined Approach for Terminological Extraction : Lexical Statistics and Linguistic Filtering*. Unit for Computer Research on the English Language. Technical Papers 5: Lancaster university.

Gledhill C. (1994). *Scientific Innovation and the Phraseology of Rhetoric; Posture, Reformulation and Collocation in Cancer Research*. Unpublished PhD Thesis. University of Aston in Birmingham.

Gledhill C. (1997). *Les Collocations et la construction du savoir scientifique*. ASp 15/18.SWALES, J. M. 1990. Genre Analysis. Cambridge : Cambridge University Press.

Harris, Z. (1968). *Mathematical Structure of Language*. New York: John Wiley and sons.

Jacquemin, C. (1996). *What is the tree that we see through the window: a linguistic approach to windowing and term variation.* Information Processing and Management, 32(4): 445-458.

McNaught, J. (1992). *Introduction to Sublanguage: A Tutorial* in Thompson, H. (ed) 1992. Record of the Workshop on Sublanguage Grammar and Lexicon Acquisition for Speech and Natural Language Processing, 7-8 January 1992, Edinburgh.

Meyer, I. & Mackintosh, K. (1996). *The Corpus from the Terminographer's Viewpoint.* International Journal of Corpus Linguistics. Vol. 1(2) : 257-285.

Moreno, M.T., J.I. Cubero, D. Berner, D. Joel, L.J. Musselman, C. Parker.(eds) (1996). *Advances in Parasitic Plant Research.* Direccion General de Investigacion Agraria. Sevilla.

Sinclair. J. M. (1996). *Preliminary Recommendations on Corpus Typology.* EAGLES. May 1996. EEC.

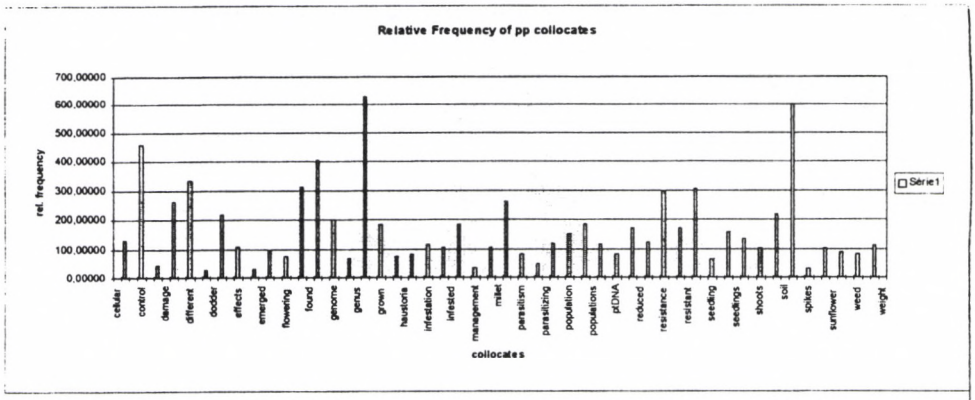
Sigiura, M. (1992). *The chloroplast genome.* Plant Molecular Biology. Vol 19: 149-168.

Swales J.M. (1990). *Genre Analysis.* Cambridge: Cambrodge University Press.

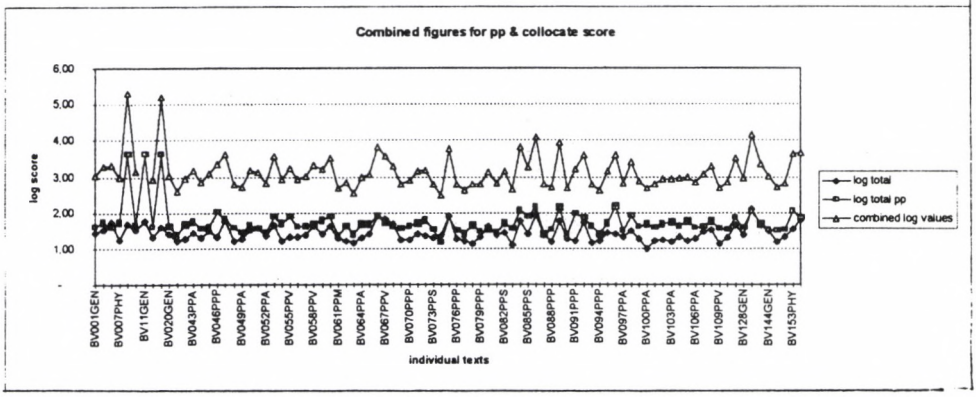
Williams G. (1998). *Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles.* International Journal of Corpus Linguistics. Vol3 (1) : 151-171.

Appendices.

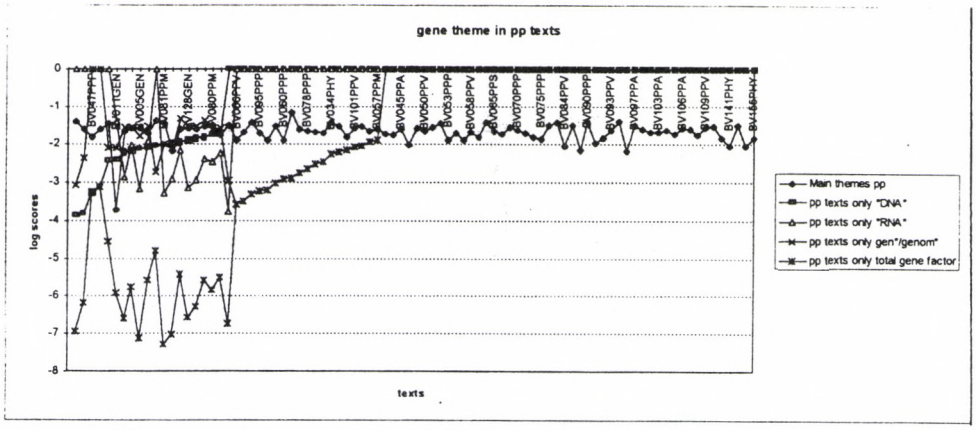
Graph 1.



Graph 2.



Graph 3.


gene theme in pp texts
log scores
texts
Main themes pp
pp texts only "DNA"
pp texts only "RNA"
pp texts only gen*/genom*
pp texts only total gene factor

Developing TEI–Conformant Lexical Databases for CEE Languages

TOMAŽ ERJAVEC – DAN TUFIŞ – TAMÁS VÁRADI

Abstract: The present paper reports on ongoing work in the INCO-COPERNICUS project CONCEDE (Consortium for Central European dictionary Encoding). The project aims to produce essential lexical databases for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The lexical databases are to be derived from existing machine readable dictionaries and marked-up in a TEI conformant encoding. Delivery of these lexica will provide one of the basic language resources for the languages in question, encoded to a common standard. We aim to make this encoding applicable to other languages and dictionary/lexical database types. The size of the LDBs developed varies by language; the average is 2500 lemmas, which provides a core vocabulary for the language. The paper gives the aims of the project, and discusses the work programme to achieve them. We highlight the issue of headword selection.

1 Introduction

For major Western European languages, there now exist a range of lexical resources, which can be used in a wide variety of contexts, e.g., as a basis for the development of language engineering applications, as training sets for machine learning methods or as material for lexicographic study. Such resources do not yet exist for most Central and East European languages.

CONCEDE , CONsortium for Central European Dictionary Encoding, is an on-going INCO-COPERNICUS project, which aims to produce essential lexical databases for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene.

Currently, the source machine readable dictionaries have been selected and acquired in digital form (they are listed in 2 below). These sources were then converted to a TEI conformant encoding, thus giving a basis for interchange and common encoding. Special attention was devoted to choosing which entries to include in the Concede LDB; we wanted to cover a broad range of phenomena that are found in dictionaries of the languages, in a manner that would be consistent and comparable across the languages.

A pilot set of 500 entries per language, encoded in TEI but with different local extensions and using different elements now serves as the common empirical basis on which to specify the LDB encoding, geared towards use in NLP. The exact definition of the LDB encoding is currently under discussion and will be reported in further work.

The rest of the paper is structured as follows: Section 2 gives an overview of the source dictionaries used in the project. Section 3 explains the methodology used in selecting the entries from the source dictionaries, which are to go into the Concede LDB. Section 4 turns to the TEI interchange encoding that the dictionaries were converted into. Section 5. gives the conclusions and the outlines the next steps in preparing the LDBs.

2 The source dictionaries

The source dictionaries for all languages concerned except Slovenian, are the standard monolingual reference dictionaries of the language. A number are still in the production process and not all entries have necessarily been written or, more often, finalised. The source dictionaries used in the project are the following:

```
Name:   Dictionary of the Bulgarian Language
Publ:   Nauka i Izkuvstvo Publishing House Sofia
Date:   1994 (4th revised edition)
Format: Microsoft Word

Name:   Dictionary of Standard Czech for School and Public
Publ:   Academia, Pragua
Date:   1994 (2nd revised edition)
Format: proprietary

Name:   Defining Dictionary of Standard Estonian
Publ:   Institute of the Estonian Language
Date:   Not yet published
Format: proprietary

Name:   Hungarian Explanatory Manual Dictionary
Publ:   Institute for Linguistic Research, Academy of Sciences
Date:   2nd edition (to be published by end of 1999)
Format: TEX/SGML

Name:   The Expanatory Dictionary of Romanian
Publ:   Univers Encyclopedic, Bucuresti
Date:   1996 (2nd edition)
Format: proprietary

Name:   English-Slovene Dictionary Oxford-DZS
Type:   bilingual, English-Slovene
Publ:   DZS publishing house, Ljubljana
Date:   Not yet published
Format: SGML, proprietary DTD
```

As can be seen, the dictionaries are in all cases quite recent, but come in a number of different formats.

3 Headword selection

One issue of common interest not only to the CONCEDE partners but, we believe, in a wider context as well is the question of finding a suitable small-scale, balanced sample of the lexicon of a language. We designed a language independent methodology which we applied to the parallel corpus "1984" by George Orwell, which was one of the deliverables in the MULTEXT-East project (Erjavec et al., 1998), and contains the same languages as the CONCEDE project. The corpus is annotated in the CESANA encoding, which is designed for linguistically analysed texts. In the "1984" case, this means that the texts are tokenised and each wordform is annotated with its lemma and grammatical information.

An empirically balanced sample per part of speech (POS) means that the POS distribution of the sample has to reflect the corresponding distribution in the corpus. A possible formula for such a distribution is given in (1). However, such a simplistic approach would have the disadvantage that it would be slanted against certain parts of speech that have few lemmas but these lemmas occur very frequently in the corpus. The number of POS lemmas chosen should not depend on the whole number of lemmas in corpus, in other words, it should not depend critically on the size of the corpus.

$$n_{POS} = \frac{N_{POS}}{N_L} * n_L \tag{1}$$

where

N_{POS}	=	number of lemmas in the corpus of a given POS
N_L	=	number of all lemmas irrespective of their part of speech
n_L	=	sample size, i.e., the number of lemmas to be chosen
n_{POS}	=	number of lemmas in the sample of a given POS

To remedy the above shortcoming, we have applied a selection method which is basically a statistical one but also relies on linguistic considerations. A total of S lemmas were chosen for all the relevant grammatical categories identified in the MULTEXT-East project, according to the frequency of their occurrence in corpus. To smoothen the selection, we divided the corpus (for each language) in NT chunks so that each of them contained exactly N_L lemmas (except for the last chunk). The selected number of lemmas for each POS is the average value over the NT chunks (2). Given that $Ni_L = N_L = S$, the formula (2) gets the simpler form shown in (3):

$$n_{POS} = \frac{\sum_{i=1}^{NT} Ni_{POS}}{\sum_{i=1}^{NT} Ni_L} * S \tag{2}$$

$$n_{POS} = \frac{\sum_{i=1}^{NT} Ni_{POS}}{NT} \tag{3}$$

Once the number of lemmas for each POS has been computed, we had to decide which lemmas would be effectively selected. We considered three frequency ranges: high, medium and low. For each POS, the sum of the high, medium and low frequency lemmas must conform to the n_POS as computed above. The frequency ranges were computed (for each POS) based on a normalised occurrence ranking of each word form. The normalised ranking of a lemma was computed as the ratio between the number of the occurrences of the respective lemma and the number of the occurrences of the most frequent lemma of that POS. Therefore the normalised ranking of a lemma is a real number less or equal to 1 (it is 1 only for the most frequent lemma(s)). For each occurrence of an inflected form of a given lemma, the respective lemma was credited with one more occurrence. The high frequency range was assigned the interval [1, 0.5], the medium frequency range the interval [0.5 0.25] and all the words with frequency range below 0.25 were considered in the low frequency range.

One would expect the first two classes (high and medium) to be scarcely populated (one, two members) and the last one overcrowded. Indeed, this would be the case if we had a Zipfian distribution. We refer here to the second of the word frequency laws (Landini, 1997) according to which if n is the frequency of a token, and N the number of words with frequency n then log(n) -0.5 log(N)

or $n * \sqrt{N}$ is constant. However, because our selection algorithm computed lemma frequencies not word frequencies, Zipf's "number-frequency" law does not apply.

Depending on the different POS, the distribution of lemmas over the three intervals was very different. For instance, in Romanian, out of the 181 nouns 10 belong to the high frequency interval, 33 to the medium frequency interval and the rest went into the low frequency interval. On the other hand, most of the functional words clustered into the high frequency interval resulting in under-populated or even empty medium and low frequency classes. The proper names, abbreviations and residuals were discarded from the selection process (usually, they are not proper items for explanatory dictionaries).

Given the structure of the corpus, as well as practical constrains in obtaining the dictionary entries in cases where the dictionary is still being produced, the number of lemmas and their percentages of the total number are given below, for the first 500 pilot sample:

Part of speech	Bg	Cs	Et	Hu	Ro	En(-Sl)
Noun	200	138	175	164	181	484
Verb	130	133	107	110	106	58
Adjective	74	75	70	100	72	122
Adverb	68	63	89	69	62	39
Open classes	472	454	441	443	421	702
Numeral	9	10	11	11	9	0
Pronoun	31	31	20	22	18	1
Conjunction	24	24	9	18	13	6
Preposition	21	21	18	9	25	11
Particle	26	2	-	-	3	-
Interjection	8	8	1	1	1	1
Article	-	-	-	3	4	-
Determiner	-	-	-	-	14	0
Closed	119	119	59	63	87	19
Total	506	506	500	506	508	721

4 Resource standardisation

To make language resources available for interchange between people, platforms and applications and to improve their 'digital longevity the benefits of standardisation have now for some time been evident. Currently, the TEI guidelines (Sperberg and Burnard, 1994) are the most detailed and widely used markup scheme. Among other text types, TEI also makes proposals concerning the encoding of dictionaries. In the first stage of the project, these were adopted as the format to up-translate to from the source data.

The conversion to TEI involves choosing a particular instantiation of the TEI, i.e., making the specific DTD (Document Type Description) to use according to which to annotated the data. The framework in which the partners worked uses the TEI.dictionaries base tagset, possibly with local extensions. In translating to this format, care was taken not to lose information from the original, hence the local extensions.

5 Conclusion

By providing much-needed lexical resources in a standard reusable way, CONCEDE is expected to further the field of language engineering in the countries concerned. By selecting words on the basis

of their frequency in naturally occurring texts for the languages, rather than by some artificial notion of which words might be useful, CONCEDE will make the lexical databases maximally useful for real applications.

The expertise developed in the project should also support general development of the language engineering industries in the countries concerned.

A further important feature of the LDBs arises from their aligned nature and relationship to the MULTEXT-East corpus. Although the lemmas for each language have been chosen independently, they will be drawn from corpora that are aligned with each other. This means that the majority of words in the LDB for one language will be correlated with words in the corresponding LDB for the other languages.

In the other papers in this volume, the work on three of the Concede language resources is discussed in more detail.

References

T. Erjavec, A. Lawson, and L. Romary, editors. *East meets West - A Compendium of Multilingual Resources*. TELRI Association, 1998.

G. Landini. Zip's Law in the Voynich Manuscript. `http://web.bham.ac.uk/G.Landini/evmt/zip.htm`, 1997.

C. M. Sperberg and L. Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford, 1994.

Markup Enhancement: Converting CEE Dictionaries into TEI, and Beyond

TOMAŽ ERJAVEC

Abstract

This paper describes the process of markup enhancement for six Central and Eastern European language dictionaries. We provide examples of the process for the English-Slovene dictionary, currently being produced by the Slovene publishing house DZS, and based on the Oxford-Hachette English-French dictionary. The TEI document type for Dictionaries is presented, followed by the process of cross-translating from original DZS SGML documents into a TEI.dictionary document and into HTML. We next discuss the development of a specialized DTD that can serve as a general model for lexical data, and provide some examples of its use.

1 Introduction

The EU project CONCEDE aims to build structured lexical databases derived from existing machine-readable dictionaries, for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The project builds on the experience and resources of the MULTEXT-EAST project [2], which developed an annotated parallel corpus for the same six languages. The CONCEDE lexical databases will be integrated with this corpus; the combined results of the two projects should constitute an integrated multilingual resource of unprecedented value.

The Text Encoding Initiative (TEI, [10] provides standard SGML-based formats for a range of text types, also dictionary data [6]. CONCEDE aims to produce lexica that are compatible with the TEI scheme. Our initial plan was to extract a subset of the TEI encoding guidelines for print dictionaries. To this end, the initial step in the project was to convert 500 entries from dictionaries in each of the project languages into the TEI.dictionary base document type. The results for each dictionary were then compared, with special attention to problems and solutions adopted in each case [8].

The initial plan within CONCEDE was to derive a subset of the TEI guidelines for dictionary encoding as the target DTD for all six lexicons. However, after examination of the encoding problems encountered in the initial 500 entries, a more radical approach was adopted. We

have designed a document type definition (DTD) based on the TEI DTD for dictionary entries that allows for the maximum of flexibility in the placement of dictionary elements, and at the same time preserves the structure of the original entry and the relations among elements implied by that structure. Because our ultimate aim is to produce lexicons suitable for use in language engineering applications, the underlying model for this DTD is based roughly on feature structures [9]. Feature structures have been heavily used in computational linguistics to model grammatical information, and their applicability to structuring lexical information has been acknowledged (see, for example, [7]). Our use of this formalism should therefore provide compatibility between the lexicons produced in the project and the structure required for their use in natural language applications.

This paper overviews the design stages of the dictionary encoding process within CONCEDE. Section 2. gives some examples from the original encoding for one of the project lexicons, the Oxford-DZS English-Slovene dictionary, in the TEI.dictionary format. With multiple conversion routines, changing data, and the source encoding already in SGML, it was essential to choose a conversion process that would exploit the advantages of standardised encoding, rather than be hindered by them. Section 3. describes the conversions in more detail. Section 4 descibes the design of the CONCEDE DTD and provides examples of its use.

2 The TEI En-Sl Dictionary Entry

The Slovene language resource for the LDB is the Oxford-DZS English-Slovene dictionary, currently in production at the publishing house DZS, and based on the Oxford-Hachette English-French dictionary [1]. The digital format of the En-Sl dictionary is SGML, with a Document Type Definition based on the markup used by Oxford University Press. The DZS markup describes structural elements, e.g. <LP> for 'lemma for phrasal verb patterns', but also rendition information, e.g. <C> for 'comma'.

Five hundred entries from the DZS dictionary were initially converted from the source SGML document type into TEI, more specifically, into the base module for dictionaries *TEI.dictionary*, producing a more structured and standardised resource. Because the DZS dictionary is still in production, the conversion to TEI also functionally validates the data, making it possible to spot and correct errors before publication.

The two direct descendants in the <BODY> of the TEI.dictionary base tagset we use are the <ENTRY> and <SUPERENTRY> elements. The former encodes entries from the original dictionary; the latter groups homonimous entries. An <ENTRY> can contain one or more syntactic or semantic <SENSE>s, where the syntactic ones can, in turn, contain their semantic senses. All three levels can contain (with certain restrictions) the form of the entry or sense, its grammatical information, translations, definitions, examples, and cross-references.

```
<entry key="beyond">
  <form><orth type='hw'>beyond</orth> <pron>bI"jQnd</pron></form>
  <sense orig='syn'>
    <gramgrp><pos>prep</pos></gramgrp>
    <sense orig='sem'>
    <trans><tr>onstran, onkraj, na drugi strani, preko</tr></trans>
    <eg orig='example'>
      <quote>beyond the city walls</quote>
      <tr>onstran mestnega obzidja</tr>
    </eg>
...
```

```
<entry key="bias ply tyre" type='compound'>
  <form>
    <orth type='hw'>bias ply tyre</orth>
    <orth type='variant'>bias ply tire</orth>
    <usg type='label'>US</usg>
    <usg type='label'>GB</usg>
  </form>
  <gramgrp><pos>n</pos></gramgrp>
  <trans>
   <usg type='label'>Aut</usg>
   <tr>diagonalni pla&scaron;&ccaron;</tr>
  </trans>
</entry>
```

The situation is complicated by various usage indicators, which are encoded as elements or (TYPE) attribute values. Verbal entries have in the original format an especially rich structure, with the markup e.g. distinguishing idioms and phrasal verb patterns. This information has been again retained either in elements (e.g. an idiom block is encoded as a <SENSE TYPE="IDIOMS">.

In our document instance we tried to retain all the information from the original format, thus making the converson from the TEI.dictionary encoding back to the original possible at least in principle. However, this policy leads to rather heavy use of the TYPE attribute.

3 Dictionary Conversion

The cross-conversion from the original to TEI.dictionary means going from documents described by one SGML DTD into those of another. We spent considerable time choosing the right tool for this job, trying to balance ease of use, expressive power and availability. The program we settled on is OmniMark LE, the 'light edtition' of OmniMark(R), available from *http://www.omnimark.com/*. While Omnimark is a commercial product, the LE incarnation is available free of charge; LE is identical to the commercial version, except for the restriction that programs cannot have more than 200 'countable actions'. So far, this not been an obstacle; the current conversion from the DZS DTD into the TEI.dictionary DTD has 44 such actions.

The conversion then proceeded in identifying, in turn, the semantics of each of the DZS elements in TEI.dictionaries and implementing the mapping. The DZS DTD defines 46 elements, some of them with quite complicated content models. The conversion therefore did not proceed only from the DTD but also took into account the actual usage of patterns in the source dictionary.

Two types of conversions were necessary; the simpler one is a context-dependent renaming of elements, as in the following Omnimark SGML translation action:

```
element GR when parent is (O2 | O3)
   output "<lbl type='gram'>%c</lbl>%n"
```

More complex are conversions that need forward reference and, in a sense, add new structure to the document. An example is the <SUPERENTRY> mentioned above: the original document marks homonims only inside the headword element, e.g. `<hw>like<hm>1</hm></hw>`. But the <SUPERENTRY> tag must be output before the start of the <ENTRY>. These cases can be solved by postponing the output until the necessary information becomes available.

We also implemented two conversions to HTML. The first is from the original, and tries to imitate the appearance of the printed dictionary, but with additional use of colors. The second is from the TEI.dictionary and formats the entry giving the descriptive names of tags, e.g. from the tag <GRAMGRP> we get 'Grammar group' in English and 'Slovnično gnezdo' in Slovene. Browsing on these formats gives, on the one hand, a feel for the finished product, and, on the other, an expanded, easy to understand standardised encoding. Both help in visualising the data and can be used as validation aids, both in the lexicographic process and in the task of LDB creation.

4 The concedeDTD

As could be expected, the initial encoding experiment for the 500 entries revealed considerable variation among the structures and elements of the CONCEDE dictionaries. Although we originally intended to use the structured TEI scheme for encoding dictionary entries, several partners used the TEI "entry free" alternative, which allows for placement of dictionary elements at any point within an <ENTRY> tag. Despite these variations, certain underlying regularities exist in all of the dictionaries, in particular, the use of a hierarchical organization that enables the factoring of information over nested levels. Although the nesting arrangement of levels in the hierarchy is not consistent across dictionaries, the use of a hierarchy to avoid re-specification of common information is virtually universal (for a discussion, see [6]).

It has been shown that all of the levels in dictionary hierarchies potentially contain the same elements [6]. There is no need, therefore, to have a proliferation of structural tags (i.e., tags marking levels in the hierarchy) which would have the same definition in the DTD. Instead, the CONCEDE DTD includes a general entry division element, <STRUC> whose name is deliberately chosen to be neutral. The <STRUC> element is used to designate structural divisions within entries, such as divisions into homographs, etc., as well as to bracket associated sets of information, most notably of three kinds: information about the "forms" of the headword (pronunciation, inflected forms, hyphenation, etc.), grammatical information (part of speech, person, number, etc.), and sense information (grouped subsenses, etc., as well as information typically associated with a given sense (usage, domain, definition, translation, etc.). Any of a set of "atomic" dictionary tags can appear within the <STRUC> element, which include orth, pron, hyph, syll, stress, pos, gen, case, number, tns, mood, usg, time, register, geo, domain, style, def, eg, etym, xr, trans, itype (see [10] for a description of these tags). For example:

```
<struc type=entry>
  <orth>demigod</orth >
  <pron>'dEmI,god<pron/>
  <pos>n</pos></struc>
  <struc type=sense>
    <struc type=subsense>
      <def>a being who is part mortal, part god.</def></struc>
    <struc type=subsense>
      <def>a lesser deity.</def></struc></struc>
  <struc type=sense>
    <def>a godlike person.</def></struc>
</struc>
```

This basic structure defines a hierarchy that can be visualized as a tree, with a node corresponding to each level. Atomic tags indicate attributes (features) associated with that node;

tag content provides the values for those features. Feature/value pairs at any node apply to all subtrees rooted at that node. Thus the encoding above can be rendered as follows:

```
[entry]  ---------|--- [sense]
orth : demigod    |       |
pron : dEmI,god   |       |---[subsense]
pos  : n          |       |    def : a being who is part...
                  |       |
                  |       |---[subsense]
                  |            def : a lesser deity
                  |
                  |--- [sense]
                       def : a godlike person
```

By traversing the tree either from the top-most node to a given terminal (or the reverse) and accumulating the information associated with each node visited during this traversal, all of the information associated with a particular *sense* of the head word (i.e., the word associated with the root node of the tree) is acquired. Thus, children of any node function as disjuncts in the feature structure formalism. We extend this formalism to allow for overriding, a frequent phenomenon in dictionary entries; in particular, when information incompatible with that given at a node higher in the tree is found (e.g., if a feature is respecified with a different value), only the information at the innermost node is retained.

The CONCEDE DTD also provides an <ALT> element that is used to designate alternates at any node. For example, the following renders a portion of the En-Sl example given in Section 2 using the CONCEDE DTD:

```
<struc type=entry key="bias ply tyre" etype='compound'>
  <orth>bias ply tyre</orth>
  <usg type='geo'>GB</usg>
  <alt>
    <orth>bias ply tire</orth>
    <usg type='geo'>US</usg>
  </alt>
  <pos>n</pos>
  <usg>Aut</usg>
  <trans>diagonalni pla&scaron;&ccaron;</trans>
</struc>
```

This corresponds to the following structure:

```
[entry]  . . . . . . . . . . . . . . [alt]
orth : bias ply tyre                  orth : bias ply tire
geo  : GB                             geo  : US
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

The dotted line indicates that [alt] is not a child of the node labelled [entry], but rather that it provides a set of alternative information. When traversing the tree to gain information about a specific use, if the [alt] information is utilized it overrides the corresponding feature/value pairs at the [entry] node. This is equivalent to providing two separate constructs:

```
[entry]
orth : bias ply tyre
geo  : GB
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

and

```
[entry]
orth : bias ply tire
geo  : US
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

We are still in the process of finalizing the CONCEDE DTD. However, this general overview provides an outline of its major features. When complete, the CONCEDE DTD will be incorporated into the Corpus Encoding Standard (CES) [4], [5].

5 Conclusion

This paper describes two stages in the development of an encoding scheme suitable for information extracted from everyday dictionaries that is intended ultimately for use in language engineering applications, in the context of the CONCEDE project. We have utilized the TEI guidelines as an intermediate DTD for encoding six Central and Eastern European language dictionaries. We have developed a DTD intended to provide the target structure for the data extracted from these dictionaries, which will render it maximally compatible with other natural language resources. To verify this possibility, CONCEDE aims to integrate its TEI dictionaries with information from an aligned and morphosyntactically annotated English original of Orwell's '1984' [3] concordancing at *http://nl2.ijs.si/corpus/* and its translations into the six project languages.

Acknowledgements

This work was supported in part by the EU project Copernicus CONCEDE.

References

[1] Marie-Hélène Corréard and Valerie Grundy, editors. *Oxford-Hachette French Dictionary.* 1994.

[2] Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, and Dan Tufiş. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada, 1998.

[3] Tomaž Erjavec and Nancy Ide. The MULTEXT-East corpus. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 971–974, Granada, 1998. ELRA. URL: http://ceres.ugr.es/ rubio/elra.html.

[4] Nancy Ide and Greg Priest-Dorman. *The Corpus Encoding Standard.* 1996. URL: http://www.cs.vassar.edu/CES/.

[5] Nancy Ide. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463-70, Granada, 1998. ELRA.

[6] Nancy Ide and Jean Véronis. *Encoding Dictionaries*, pages 167–180. Kluwer Academic Publishers, Dordrecht, 1995.

[7] Nancy Ide, Jacques Le Maitre, and Jean Véronis. Outline of a Model for Lexical Databases. *Current Issues in Computational Linguistics: In Honour of Don Walker. Linguistica Computazionale* IX, X, pages 283-320, Pisa, 1995. [reprinted from *Information Processing and Management*, 29, 2, pages 159-186]

[8] Adam Kilgarriff, editor. *Concede Deliverable 2.2: Dictionary Encoding Schemes.* University of Brighton, 1999.

[9] Stuart Shieber. *An Introduction to Unification-based Approaches to Grammar.* CSLI Lecture Notes Series, University of Chicago Press, Chicago, 1986.

[10] C. M. Sperberg-McQueen and Lou Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford, 1994.

TEI–Encoding of a Core Explanatory Dictionary of Romanian

DAN TUFIŞ – GEORGIANA ROTARIU – ANA-MARIA BARBU

ABSTRACT

The efforts on development of large Lexical DataBases (LDB) are just emerging in most of the CE-countries. CONCEDE is an EU project aiming at harmonising the methodologies, tools and (to a less extent) resources for building LDBs for six CE-languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene.

The paper addresses the specific problems concerning the process of TEI encoding of the Romanian Explanatory Dictionary. The Romanian Explanatory Dictionary (DEX, second edition, 1996) is the reference dictionary of Romanian and it was developed by the Institute for Linguistics of the Romanian Academy and published by the Univers Enciclopedic Publishing House. DEX is meant for a wide public and therefore the lexicographic content is relatively rich with head-words belonging to what is generally called a basic vocabulary for one language, including regional variants, but containing also various technical or specialised terms as well as various neologisms and recent lexical imports.

1. Introduction

The efforts on development of large Lexical DataBases (LDB) are just emerging in most of the CE-countries. CONCEDE is an EU project aiming at harmonising the methodologies, tools and (to a less extent) resources for building LDBs for six CE-languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The project adopted an incremental approach, therefore having a generic sampling method for deciding at each step on what headwords to include into the lexical database was important. This procedure is described in the paper that gives an overall presentation of the CONCEDE project. Here, we will address the specific problems concerning the process of TEI encoding of the Romanian Explanatory Dictionary.

2. The Romanian Explanatory Dictionary

The Romanian Explanatory Dictionary (DEX, 1996) is the reference dictionary of Romanian and it was developed by the Institute for Linguistics of the Romanian Academy and published by the Univers Enciclopedic Publishing House. DEX is meant for a wide public and therefore the lexicographic content is relatively rich with head-words belonging to what is generally called a basic vocabulary for one language, including regional variants, but containing also various technical or specialised terms as well as various neologisms and recent lexical imports. It contains about 65,000 entries each of them containing plenty of information, some of it in an explicit way, some other in the implicit format (layout conventions). The information categories are the following: head-word, accentuation, inflected forms, accent shift (where the case), pronunciation, grammatical information on the head-word and inflected forms, sense definitions, references to other head-words or other sense definitions, phrasal constructions, usage information, lexical relations (synonyms, hypernyms, hyponyms and sometimes antonyms), headword variants, etymology. In most cases, examples and synonymy series accompany sense definitions. Whenever the case, and distinctly marked, the dictionary entries contain expressions and locutions headed by the head-word (or one of its inflected forms).

Given that the copyright for the electronic version of the dictionary was not in the hands of the Romanian Academy and the copyright holders did not agree to provide the raw texts, we had to keyboard the information provided in the printed dictionary. When doing so, all the information implicit in the layout (see section 3.1.3) was made explicit by means of specific mark-up.

Because within the CONCEDE project we had not enough resources to keyboard all the entries and trying to give as much potential as possible to our work for various applications, we decided to type in selectively the most frequent words we found in our various corpora (more than 10.000.000 words). We extracted a list of about 15.000 candidate lemmas existing in DEX, out of which more than 10.000 entries were already keyboarded. According to preliminary tests made on the annotated part of our corpora containing about 1.000.000 words annotated in conformance with CES dtd (Ide, 1998, Dimitrova & all, 1998, Ide, Veronis, 1995, Ide, Veronis, 1994), these 10.000 lemmas (and their inflected forms) will cover at least 90% of new texts. Therefore, encoding these lemmas into a LDB would create a useful lexical resource for most of NLP applications.

3. Extraction and conversion of data from the printed dictionary

3.1. The extraction process

Professional typists ensured the keyboarding process and they were instructed to follow exactly the layout of the printed dictionary, except for some well-specified conditions (see below).

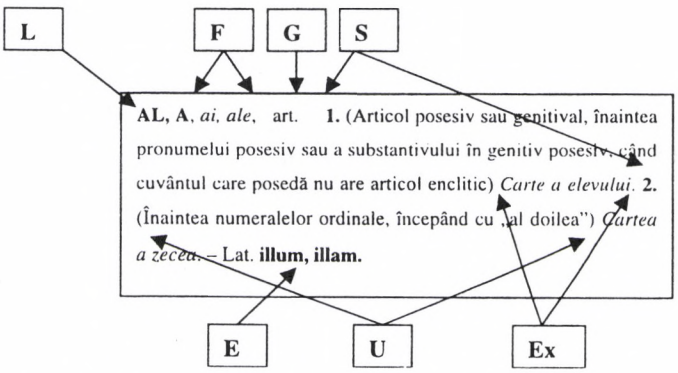
We developed a program that is aware of all the conventions in the printed form of DEX (described in the sections 3.1.2 and 3.1.3) as well about the TEI.dictionary. This program is rather sensitive to a pre-specified order of the information types but is quite permissive in defining the lexicographic conventions.

One should notice that the current version of the program is not able to achieve a fully automated conversion (the average is about 80% and although we are confident that this figure could be improved we are rather doubtful concerning a fully error-free automated procedure). This is due to the fact that there are entries that for one reason or another do not conform to the general structuring and conventions mentioned before (and discussed next) and on the other hand, and this is a more serious reason, certain encoding decisions do not depend on syntactic criteria, but require interpretation and human judgement.

For the vast majority of the selected entries, the conventions were strictly observed and this fact significantly simplified the automatic conversion from the MSWord format (actually the Word files were exported as HTML files and the conversion started from this format) to the target SGML encoding.

3.2. The ordering of various information types in the printed dictionary

Information associated to a head-word in DEX observes the structuring and typographical conventions as shown in Figure 1 and explained below.



L - lemma (head-word); **F** - inflected forms; **G** - grammatical information; **H** - homographs; **S** - sense; **Ss** - secundary sense; **D** - definition; **V** - variants; **E** - etymology; **Ex** - examples; **U** - phrasal unit **I** - usage information

Figure 1: The layout of an entry in DEX

The order of different classes of information is (more often than not) as follows:

1. The lemma form of the head-word
2. Inflected forms, where the case; when the inflected forms have specific grammatical information this is specified together with the corresponding inflected form. If the inflected form is associated with a specific sense of the head-word, then the inflected form is accompanied by a reference to the specific sense;
3. Global grammatical information
4. The explanation(s) of the head-word; they are either direct definitions (grouped on various senses) or indirect definitions (specified as references to other head-words or in case of functional words by explaining their usage)
5. Information on the pronunciation, variants, irregular inflectional paradigms
6. Etymological information

3.3. Lexicographic conventions

Thanks to the systematic usage of some prescribed lexicographic conventions (some of them explicitly documented in the preface of the printed dictionary, others used implicitly but consistently) the conversion from HTML format to the target encoding was substantially simplified. This section briefly reviews these conventions.

- The head-words are always written in uppercase characters. If they belong to a homonymy class, the head-words (or the words appearing in the definitions of some head-words) are differentiated by numeric superscripts. The stressed vowel is always represented in the printed dictionary by an accented letter (á, é, ó, etc.) In the keyboarded format of the printed dictionary, the accented vowels in the head-words were represented as a quote followed by the corresponding vowel ('a, 'e, 'o, etc.).
- The distant senses are marked by uppercase letters (A, B, etc) or Roman numeral (I, II, etc.). The explicit related senses are numbered by means of lowercase letters (a, b, etc.) or Arab numbers (1, 2, etc.). The senses that are dependent on a main sense are marked by a black diamond (♦). The phrasal units (locutions, expressions, compounds, etc.) subordinated to a main sense, as well as some specialised senses which do not require a new definition, are signaled by an white diamond (◊).
- The equal sign (=) signals the definition of a phrasal unit. Double quotes surround a collocation if they appear inside a sense definition or a gloss if they appear in the etymological section.
- The square brackets contain pronunciation information, lexical variants, specific irregular inflected forms. Each type of information marked by square brackets is differentiated by specialised labels (Pr: - for pronunciation, Var: - for variants, or a grammatical label, such as Prez.ind., for irregular inflected forms). In case that more than one type of squared information is provided, they are separated by dashes.
- Usage information is provided between parenthesis.
- The symbols <, > and + are used in providing further information on the etymology of the head-word. The etymological information always appears as the last field of the entry and is systematically introduced by a dash.

4. TEI-encoding of the dictionary

The first experimental step towards encoding a sample of DEX entries, tried to preserve both possible views: the editorial and the lexical views (for a detailed discussion on possible views on a paper dictionary and on methods to conciliate them, see Ide, Veronis, 1995). With the variety of information existing in DEX (but also due to some inconsistencies that exist in the

printed dictionary), ensuring TEI-conformance and preserving the editorial and lexical views proved to be practically impossible. Thus, we reformulated our goal so that to ensure as much TEI conformance as possible and to fully preserve the lexical information in the printed dictionary.

Once we finally decided to adopt the lexical encoding schema we committed ourselves to a prescribed order of the elements inside the <entry> element as shown in Figure 2.

```
<entry>
        <form>.................</form>
        <gramGrp>...........</gramGrp>
        (<sense>.............</sense> | <xr>...........</xr>)+
        (<etym>...............</etym>)*
</entry>
```

Figure 2: The structure of an encoded entry

In most cases the ordering and structure in Figure 2 is observed by the printed dictionary (see Figure 1) but whenever this was not the case, the necessary positional changes where done so that to comply with the prescribed structure. Additionally, we got rid of the special graphical characters which were either irrelevant for our purposes or became redundant due to the explicit SGML mark-up. The phrasal units and definitions where expanded wherever the case so that both the clarity and ease of exploitation improved (as shown by the experiments we made by using SGML-QL (Véronis, 1997)). The preliminary experiments with storing and exploitation of our TEI dictionary as a standard database (ORACLE) show that the automatic conversion is much more feasible.

This initial encoding proved to be a challenge for various reasons:

- the entries have not always homogeneous structure; to overcome the observed inconsistencies, initially, we used the *entryFree* element, the consistency restrictions of which are very loose, thus easily allowing an intermediary straightforward encoding of all lexical information provided in the printed dictionary. The displaced information (with respect to our ordering) was easily spotted and moved in the appropriate position.
- the lexical information is frequently implicitly specified. For instance, consider the lexical entry below:

  CENTR'AL, -Ă, *centrali, -e,* adj., s.f. ....

  The grammatical information s.f. (feminine noun) refers not to the lemma, but to the word-form obtained by adding the suffix **-Ă**, which probably is not easy to infer for a non-native speaker of Romanian. The proper expansion would associate **CENTR'AL,** *centrali* with the implicit information "adj.m." (masculine adjective) and **CENTR'ALĂ,** *centrale* with "adj.f., s.f" (feminine adjective or feminine noun).
- another problem we were faced with was that a large part of information in the printed dictionary is provided in a format which was meant to save space (relying on the human reader ability to expand the compressed form). Saving space was done in basically two ways: the first one, posing no problems to machine processing, concerns the abbreviations used throughout the dictionary; the second one, trouble making, concerns the phrases which are printed in a shortened form (definitions, examples, collocations, etc.). Such a phrase consists of fix and variable components, but unfortunately expanding the meaningful combinations relies, as said before, on the human reader of the dictionary. For instance a simple case of shortening an expression is the following: "*în* (sau *din) două vorbe* (sau *cuvinte)*". It stands for the following four variants: "*în două vorbe*", "*din două vorbe*", "*în două cuvinte*", "*din două cuvinte*". For the sake of consistency (and conformance with TEI-

encoding lexical view recommendations) we finally decided to expand both the abbreviations and the shortened phrases.

The intermediary encoding mentioned before, paved the way for migrating from *entryFree* to the more constrained *entry* element. However, the basic structure of the *entry* element of *tei2dict.dtd* had to be slightly modified for covering all lexical information one finds in DEX (e.g. the content model of the *gramGrp* element, *lbl* domain, etc.). In the figures below, there are described the main extensions we made to the basic dtd, in order to accommodate all available information as provided in the printed dictionary (the extension files below are appropriately referred to in the tei2.dtd).

```
<!-- Suppressing elements that are modified in the          -->
<!--         entity tei.extension.dtd                       -->
<!ENTITY % usg 'IGNORE'      >
<!ENTITY % gramGrp 'IGNORE' >
<!ENTITY % def 'IGNORE' >
<!ENTITY % form 'IGNORE' >
<!ENTITY % sense 'IGNORE' >
```

Figure 3: concede.ent

```
<!-- The following declarations define revised tags -->
<!ELEMENT %n.usg;          - O  (%paraContent; | %n.colloc; | %n.lbl; )+  >
<!ATTLIST %n.usg;               %a.global;
                                %a.dictionaries;
          type                  CDATA                #IMPLIED
          TEIform               CDATA                'usg'                  >
<!ELEMENT %n.gramGrp;      - -  (%m.gramInfo | %m.morphInfo | %paraContent)* >
<!ATTLIST %n.gramGrp;           %a.global;
                                %a.dictionaries;
          TEIform               CDATA                'gramGrp'              >
<!ELEMENT %n.def;          - O  (%paraContent;)                 >
<!ATTLIST %n.def;               %a.global;
                                %a.dictionaries;
          type                  CDATA                #IMPLIED
          TEIform               CDATA                'def'          >
<!ELEMENT %n.form;         - -  (%m.formInfo | %n.stress | %n.gramGrp |
%paraContent)+ >
<!ATTLIST %n.form;              %a.global;
                                %a.dictionaries;
          type                  CDATA                #IMPLIED
          TEIform               CDATA                'form'         >
<!ELEMENT %n.sense;        - -  (%n.sense; | %m.dictionaryTopLevel
                                | %m.phrase | %n.lbl | #PCDATA)*           >
<!ATTLIST %n.sense;             %a.global;
                                %a.dictionaries;
          level                 NUMBER               #IMPLIED
          TEIform               CDATA                'sense'        >
```

Figure 4: concede.dtd

5. Adapting teidict2.dtd for DEX

Up to this phase of the project we avoided any lexical loss in the SGML encoding as compared to the (implicit or explicit) information provided by DEX. Therefore, we slightly modified teidict2.dtd.

As said before, the initial encoding used <entryFree> markup, but following the ordering in Figure 2. We encoded about 300 entries out of the 500 selected as described in section 2.

The modification of the <entryFree> markup for <entry> was done by using XEMACS macros and appropriate specifications (see Figures 3 and 4) in the extension files TEI.extensions.ent and TEI.extensions.dtd of the elements and attributes which were in the initial encoding (and we wanted to preserve).

In the following we will dwell on these modifications and exemplify why they were needed.

1. The <stress> element is included into the <form> element and because in written Romanian the accent is not marked, <stress> and <orth> information was kept distinct.
2. Another element that was embedded into <form> is <gramgrp>. This was necessary in order to make possible to associate grammatical information that was pertinent only to a specific inflected form or variant. For instance the inflected forms in direct cases (Nominative and Accusative) both singular or plural are implicit, but for the oblique cases the orthographic forms are explicitly associated with case information.

Exemple 1

```
<form>
      <form>
            <orth>acel</orth>
             <stress>ac`el</stress>
      </form>
      <form type="inflected">
            <form>
                  <orth>acea</orth>
                  <stress>ace`a</stress>
            </form>
            <form>
                  <orth>acei</orth>
                  <orth>acele</orth>
            </form>
            <form>
                  <orth>acelui</orth>
                  <orth>acelei</orth>
                  <gramgrp>
                        <case>genitiv, dativ</case>
                        <number>singular</number>
                  </gramgrp>
            </form>
            <form>
                  <orth>acelor</orth>
                  <gramgrp>
                        <case>genitiv, dativ</case>
                        <number>plural</number>
                  </gramgrp>
            </form>
      </form> ...
</form>
```

3. Romanian is a strongly inflected language and therefore preserving the morphological information provided in DEX is absolutely necessary (when this information is provided, usually the wordform in case is irregular). Therefore the entity %m.morphInfo was included into the content of <gramGrp>.

The morphological information is systematically provided for the head-word, but this is also true for the inflected forms that have variants. In such cases, the morphological information is provided at the beginning of the entry.
When the inflectional paradigm of the head-word is irregular, DEX specifies the irregular forms together with the corresponding morphological towards the end of the entry (just before the etymological information). According to the encoding schema in Figure 2, these wordforms and associated grammatical information were moved into the first <form> element which contains the orthographic, orthoepic and morphological information applying for the head-word.

It also happens that specific morphological information may apply just for some senses (see example 2) or phrasal units (see example 3).

Exemple 2

```
<form>
      <orth>fi</orth> ...
</form>
<gramgrp>
      <pos>verb</pos>
      <itype>conjugarea IV</itype>
      <subc>intranzitiv</subc>
</gramgrp>
<sense n="A">A.
      <gramgrp>
            <pos>Verb predicativ</pos>
      </gramgrp> ...
  </sense>
```

Exemple 3

```
<sense type="phrase">
      <gramgrp>
            <pos>locuţiune verbală</pos>
      </gramgrp>
      <re type="loc">
            <form>
                  <orth>A-i fi cuiva drag (cineva sau ceva)</orth>
            </form>
            <sense>
                  <def>a-i plăcea, a îndrăgi, a iubi. </def>
            </sense>
      </re>
</sense>
```

The HTML-SGML conversion program we mentioned before, explicitly generates all the implicit grammatical information of the head-words (infinitive for verbs, masculine singular for adjectives or pronouns, etc.)

4. In order to easily identify the words that collocate with a specific head-word we included the <colloc> element into <usg> (see example 4).

Exemple 4

```
<entry type="hom" n="1">
      <form>
```

```
            <orth>an</orth>
      </form>
      <gramgrp>
            <pos>adverb</pos>
      </gramgrp>...
        <sense>
                <usg type="style">figurat</usg>
                <usg type="colloc">Precedat de <colloc>mai</colloc></usg>
                <def>Acum câţiva ani.</def>
        </sense>...
</entry>
```

6. Related entries

The <re> markup was used for encoding the phrasal units. DEX contains many phrasal units and distinguishes among locutions, expressions, syntagms, constructs, compounds. The description of a phrasal unit exhibits a similar structure (but simplified) to the one of a regular entry. A maximal description of a phrasal unit contains all the toplevel elements shown in Figure 2, except for the <etym> and <xr>.

Below we provide examples for each type of phrasal units mentioned before.

Exemple 5

```
<sense type="phrase">
      <re type="expr">
            <form>
                  <orth>de la a la z</orth>
            </form>
            <sense>
                  <def>de la început până la sfârşit</def>
                  <def>totul, în întregime.</def>
            </sense>
      </re>
</sense>
```

Exemple 6

```
<sense type="phrase">
      <gramgrp>
            <pos>locuţiune adverbială</pos>
      </gramgrp>
      <re type="loc">
            <form>
                  <orth>An de an</orth>
                  <orth>An cu an</orth>
            </form>
            <sense>
                  <def>în fiecare an, mereu. </def>
            </sense>
      </re>
</sense>
```

Exemple 7

```
<sense type="phrase">
      <re type="compound">
            <form>
```

```
                <orth>cinci-degete</orth>
            </form>
            <gramgrp>
                <pos>substantiv</pos>
            </gramgrp>
            <sense>
      <def>plantă erbacee târâtoare, cu frunzele formate din cinci
      foliole şi cu flori galbene <term lang="LA">(Potentilla
      reptans)</term>. </def>
            </sense>
       </re>
</sense>
```

Exemple 8

```
<re type="synt">
      <form>
            <orth>Faţă de masă</orth>
      </form>
      <sense>
            <def>material textil, plastic etc. folosit spre a acoperi o masă
                (când se mănâncă sau ca ornament). </def>
      </sense>
</re>
```

Acknowledgements

The work reported here was jointly funded by the CONCEDE European project (PL96-1142) and by a grant of the Romanian Academy (GAR#187/1998)

References

DEX, (1996). Coteanu, I., Seche, L., Seche, M. (coord.): Dicţionarul Explicativ al Limbii Române, Ediţia a II-a, *Univers Enciclopedic*, Bucureşti, 1996

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, J. H., Petkevič, V., and Tufiş, D. (1998): Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages" *In Proceedings of COLING-ACL'98,* Montreal, Canada, 315-319

Erjavec, T., Ide, N. (1998): The Multext-EAST Corpus. In *Proceedings of the First International Conference on Language Resources and Evaluation,* LREC'98, Granada, 1998, pp. 971-974

Ide, N., Véronis, J., eds. (1995): Text Encoding Initiative. *Kluwer Academic Publishers*, Dodrecht / Boston / London, 1995

Landini, Gabriel (1997): Zipf's laws in the Voynich Manuscript. *http://web.bham.ac.uk/G.Landini/evmt/zipf.htm*

Ide, N. (1998) Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora In *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain. See also http://www.cs.vassar.edu/CES/.

Ide, N. and Véronis, J. (1994): Multext (Multilingual Tools and Corpora). In *Proceedings of the 14th International Conference on Computational Linguistics*, COLING'94, Kyoto, pp. 90-96.

TEI Encoding of the Hungarian Explanatory Manual Dictionary

CSABA ORAVECZ – TAMÁS VÁRADI

Abstract: The paper presents findings of ongoing work to convert the Hungarian Explanatory Manual Dictionary to TEI compatible encoding as part of the CONCEDEproject. It discusses the need to restrict and extend the TEI guidelines to accomodate the special Hungarian lexicographic requirements. The proposed modifications to the teidict2.dtd are presented in detail. The paper highlights a number of encoding problems which arise from the difficulty to reconcile the lexical view of the dictionary with the editorial view. Conversion between the two representations are sometimes far from trivial and require manual intervention.

1 Introduction

The present paper reports on work carried out as the Hungarian contribution in the CONCEDE project (for an overview of the project see Erjavec et al., 1999). The paper is structured as follows: Section 2 describes the source data for the projected Hungarian lexical database, Section 3 introduces the encoding scheme adopted, Section 4 lists the proposed modifications to the teidict2.td and Section 6 carries some conclusions.

2 Description of the source data

The Hungarian Explanatory Manual Dictionary (HEMD) (Juhász et al., 1972) is a major academic work first published in 1972. It was intended to serve as an updated, more practical compendium to the seven volume Hungarian Explanatory Dictionary, which was completed in the 1950's. It is a general purpose comprehensive manual dictionary compiled by a team of lexicographers at the Institute for Linguistic Research of the Hungarian Academy of Sciences. The dictionary consists of cc. 70,000 headwords chiefly aimed at covering standard literary and colloqual Hungarian but also including a wide range of regiters which are thought to be part of standard Hungarian.

The entries typically have a fairly elaborate structure breaking down into the following fields: head-word, part of speech, grammatical information, sense definition, usage information, examples, phrasal constructions (including collocations, sayings, idomatic expressions, each with their own

examples, usage and sense definitions), cross-references to other head-words, and etymologies. As Hungarian pronunciation is thought to be highly regular and generally presenting no problem to native speakers, pronunciation is only mentioned sporadically. The same strategy is followed with respect to grammatical information. Morphological information about how headwords are inflected are not given even for irregular cases. Instead, a concise table of verb and noun paradigms is provided in the introduction.

Both for its professional merits (depth and breadth of lexical coverage) as well as practical issues such as copyright and availability, the HEMD was chosen for conversion into a lexical database.

The HEMD is currently being revised, with the 2nd edition due to be completed by the end of 1999. The opportunity of the revision was used also to develop an electronic version of the material. However, lack of time and the lexicographers' preference for traditional paper-and-pen methodology resulted in the fact that the revision work and the transfer into machine readable form were carried out as separate processes. The revised version was prepared on paper and was handed over to a colleague for encoding. Fortunately, the electronic form of the revised version was planned with sufficent care. A provisional DTD was compiled and the text of the dictionary was to be encoded in SGML using the SGML editor WriterStation. It was planned that the DTD was to be regularly evaluated and when necessary modified to accomodate unforseen cases.

In the event, however, the encoding work was carried out in TEX through a suite of TEX macros. They were devised to be so close to the SGML format that two way conversion between TEX and SGML annotation proved a trivial task with the help of a file editor. Unfortunately, the envisaged regular updates of the provisional DTD have become less and less frequent until they have ceased to take place. Because the encoding in TEX did not involve any periodic checks on the validity of the notation, as would have been the case with the use of a proper SGML editor, the compliance of the data to the DTD became increasingly slack. It was decided, therefore, that rather than attempting to patch the provisional DTD, a new DTD would be applied with the data annotation scheme modified as necessary.

3 Selection of an encoding scheme

As one of the basic boals of the CONCEDE project is to devise and validate a TEI-conformant encoding standard, an obvious choice of an initial formalism was to consider the guidelines of the TEI Dictionary Working Group. However, the price of wide converage and generality of TEI is that it is unnecessarily large and not restrictive enough for a particular domain. Accordingly, the corresponding subset of the TEI DTD, the teidict2.dtd and the teidict2.ent entity set was adopted as a base from which to develop an encoding scheme more tailored to the data at hand. Because the dictionary data was already well structured and supplied with markup, we decided to use the <entry> elements instead of the looser <ENTRYFREE> type. Even so, it was found that the definition of <ENTRY> needed both restriction and extension at the same time.

4 Proposed modifications to the teidict2.dtd

4.1 Introducing the <expr> and <ExprGrp> elements

A frequent feature in HEMD is the use of longer phrases, sentences that are adduced to throw light on how the entry word is used in context. Such items may come with their own usage and style labels, definitions and sometimes etymology. HEMD distinguishes three types among them i.e. sayings, pseudo-sayings, and similes. The element <EXPRGRP> serves to contain several <EXPR> elements, which share their <TYPE> attribute. They have a more elaborate structure than what could

be encoded within <USG type="coll"> tags, on the other hand, they are more constrained than <RE> as they do not feature on the same level in the hierarchy as <ENTRY>. Hence we felt the need to set up a separate element to deal with this material.

Example:
ablak ...
Sz: *gúny: nem teszi (ki az) az -(á)ba:* ezzel ugyan nem fog dicsekedni!; *az -on dobja v. hajítja v. szórja ki a pénzt:* fölöslegesen költekezik.
window ...
Sz: *sarcastic: does not put (out) in (the/his) -:* this is certainly nothing to show off with; *throws/casts/hurls money through the -:* spends money unnecessarily.

```
<exprGrp type="SZ">
 <expr><usg>gúny:</usg>
   <eg rend="POST colon">
   <q>nem teszi <hint>ki az</hint> <oref><hint>á</hint>ba</q>
   </eg>
   <def>ezzel ugyan nem fog dicsekedni!;</def>
 </expr>
 <expr>
   <eg rend="POST colon">
   <q>az <oref>on dobja <lbl>vagy</lbl> hajítja
      <lbl>vagy</lbl> szórja ki a pénzt:</q>
   </eg>
   <def>fölöslegesen költekezik.</def>
 </expr>
</exprGrp>
```

4.2 Extending the content model or attribute list of some tags

4.2.1 Adding <usg> and <lbl> to the content model of the <def> tag

ad ...**3.**
<Ruhadarabot> vkinek a testére húz, segít.
give ...**3.**
<Piece of clothes> sy his body pull, help.

```
<def>
 <usg type="hint">Ruhadarabot</usg> vkinek a testére húz, segít.
</def>
```

alatt ...**8.**
vmely feltétel, ürügy ~ : (helyesen:) feltétellel, ürüggyel
under ... **8.**
some condition, pretext ~: (correctly:) under some condition, pretext

```
<def>
 <usg type="acc">
```

```
   <lbl>helyesen</lbl> feltétellel, ürüggyel
 </usg>.
</def>
```

4.2.2 Adding <type> attirubute to the <def> and <pos> tags

In Hungarian lexicographic tradition, sometimes a definition is given in terms of describing the situation when the particular word or phrase is used. In such cases, the <DEF> is marked with a <TYPE> attribute to indicate that this is not a genuine definition.

a ... **3.** ... <A figyelmet vmely hangra irányító szóként.>
a ... **3.** ... <Used as a word to draw attention to a sound.>

```
<def type="rep">A figyelmet vmely hangra irányító szóként.</def>
```

Some suffixed forms are defined in <DEF> tags of similar attribute by explicating the composition of the form concerned. E.g.

ahhoz ... Az *az* mutnévm ragos alakja
ahhoz ... Inflected form of the demonstrative pronoun *az*

```
<entry><lemma>ahhoz</lemma><sense><def type="rep">Az <mention>az
</mention>mutnévm ragos alakja</def></sense>
```

Sometimes the <POS> tag carries a similar indirect definition.
de- (idegen szavak előképzője)
de- (prefix in foreign words)

```
<gramgrp>
  <pos type="rep">idegen szavak előképzője</pos>
</gramgrp>
```

4.2.3 Restricting the contents of the <eg> element

In the teidict2.dtd the element <%N.EG> included <%M.DICTIONARYPARTS>. It was allowed to recur in any position and could introduce a whole bag of unnecessary elements. It almost looks like a bug in the original design and was eliminated in our DTD.

```
<!ELEMENT %n.eg; - O          (%n.q; | %n.quote; | %n.cit;)+
                    +(%m.dictionaryParts | %m.formPointers)>
```

4.3 The revised dtd

Technically, the modifications were implemented in comformity with the TEI guidelines as described in Chapter 29 of the TEI manual (Sperberg and Burnard, 1994). In conclusion we display the proposed modifications to the teidict2.dtd in full.

```
<!--          concede.dtd  -->
<!-- The declarations below define new extensions -->
<!ELEMENT %n.exprGrp     - -  (%n.expr;)+                          >
<!ATTLIST %n.exprGrp;         %a.global;
                              %a.dictionaries;
          type                (space|Sz|Szh|Szj)   'space'        >

<!ELEMENT %n.expr;       - -  (%n.q | %n.usg; | %n.lbl; | %n.def;
                              | %n.eg; | %n.etym;)*               >
<!ATTLIST %n.expr;            %a.global;
                              %a.dictionaries;                    >

<!-- The following declarations define revised tags -->

<!ELEMENT %n.def;        - O  (%paraContent | %n.usg; | %n.lbl;)*>
<!ATTLIST %n.def;             %a.global;
                              %a.dictionaries;
          TEIform             CDATA                'def'
          type                CDATA                #IMPLIED       >

<!ELEMENT %n.pos;        - O  (%paraContent;)                      >
<!ATTLIST %n.pos;             %a.global;
                              %a.dictionaries;
          TEIform             CDATA                'pos'
          type                CDATA                #IMPLIED       >

<!ELEMENT %n.usg;        - O  (%paraContent | %n.usg)*             >
<!ATTLIST %n.usg;             %a.global;
                              %a.dictionaries;
          TEIform             CDATA                'usg'
          type                CDATA                #IMPLIED       >

<!ELEMENT %n.eg;         - O  (%n.q; | %n.quote; | %n.cit; | %n.usg)+

                                                   +(%m.formPointers)
                                                                  >
<!ATTLIST %n.eg;              %a.global;
                              %a.dictionaries;
          TEIform             CDATA                'eg'          >
```

5 Encoding problems

In a few cases the encoding presented some difficulty not because of the lack of notation devices but on the contrary because TEI offered various alternative ways to encode the same phenomenon. We set out below the particular solutions we adopted.

1. Pointers and cross referencing
   We have decided to implement the cross-reference to a numbered homonym by forming a

unique id of the homonym out of the orthographic form and the homonym number and include as an attribute of the <FORM> tag. The hardwired number is local to the entry concerned and is independent of any id number the entry itself may assume. This arrangement is expected to make any possible updating of the cross-references fairly straightforward.

áll[1] ...
stand[1] ...

```
<entry type="hom" n='1'>
   <form id='áll.1' type="lemma"><orth>áll</orth></form>
```

állandó ...
[←*áll*[1]]
standard ...
[←*stand*[1]]

```
<etym><mentioned><ptr target='áll.1'></mentioned></etym>
```

2. Optional and alternate forms

In order to save space, the HEMD frequently resorts to putting optional information in parentheses and to merging alternative forms into a single disjunctive expression.

- optionality

ablak ... **1.** Épületen, járművön a világosság és a levegő bebocsátására való (zárható) nyílás. **window 1.** On a building or a vehicle, a (closable) hole to let the light and air in.

```
<def next=ablak.1.3 id=ablak.1.2>
 Épületen, járművön a világosság és a
 levegő bebocsátására való </def>
<def prev=ablak.1.2 opt=y next=ablak.1.4 id=ablak.1.3>zárható</def>
<def prev=ablak.1.3 id=ablak.1.4> nyílás.</def>
```

- compressing alternate forms into a single expression
  While it may be argued that the optional information in the above example may be left unencoded as part of the text of the definition (at least as a first approximation), the alternatives in the example below need to be spelt out in order to derive the explicite forms the expressions take. Sometimes the generation of all possible alternatives requires manual intervention as not all the resulting forms are well formed or because it is difficult to establish how to spell out the compressed expression.

ad ... **11.** ... *Vmire v. vminek -ja magát v. a fejét:* vmire szánja, ill. vminek átengedi magát
give ... **11.** ... *To sth o. as sth -oneself o. one's head:* gives himself in to sth

```
<exprGrp><expr>
 <eg rend="POST colon">
  <q>Vmire <oref>ja magát</q>
  <q>Vmire <oref>ja a fejét</q>
  <q>Vminek <oref>ja magát</q>
  <q>Vminek <oref>ja a fejét</q>
 </eg>
 <def>vmire szánja, <lbl>ill</lbl> vminek átengedi magát.</def>
</expr></exprGrp>
```

ad ... **14.** ... *áldását -ja:* a) *vkire:* megáldja; b) *vmire: átv is* jóváhagyja;
give ... **14.** ... *one's bless -:* a) *sy:* bless; b) *sth: fig too* concede;

```
<exprGrp><expr>
 <eg rend="POST colon">
  <q>áldását <oref>ja vkire:</q>
 </eg>
 <def n="a">megáldja;</def>
 <eg rend=''POST colon''>
  <q>áldását <oref>ja vmire:</q>
 </eg>
 <usg>átv is</usg>
 <def n="b">jóváhagyja;</def>
</expr></exprGrp>
```

3. Abbreviated keywords in definitions and examples

Another frequent space saving device is the abbreviation of the headword to its intial letter when it is used in the body of the definition.

ablak ...
-nyílás ... Épület falában ablaknak, ill. ablak helyett hagyott ny.
window ...
-opening ... In the wall of buildings an o. for or instead of a window.

```
<def>
 Épület falában ablaknak,
 <lbl>ill</lbl> ablak helyett hagyott
 <abbr orig="ny.">nyílás</abbr>
</def>
```

6 Conclusions

Our work so far has suggested that preparing a lexical database from a machine readable version of a paper edition of the dictionary is far from being a trivial task. Even when the source material is available in some sort of annotated form, it is likely to be too closely following the space saving notational expediences that dominate the content and layout of the paper edition. Our source dictionary in particular is very rich in such compression techniques (e.g., abbreviations, parenthetical and disjunctive expressions etc.).

For a strictly lexical (as against typographical) view of the dictionary, it is sometimes difficult to tease out the component that inherently carry content from those that serve are dictated by space constraints. Should disjunctive phrases within the body of definitions be resolved, for example? The problem is compounded by the fact that often the interpretation of these space saving notational devices require a degree of human intelligence that is difficult to automate. Furthermore, in a number of cases the mechanical explication of alternatives lead to non-existent or ill-formed expressions, a situation which again calls for manual editing.

The conflict between the typographical and the lexical views of the dictionary that we find particularly acute in the case of HEMD, may be eased by the introduction of levels of granularity of encoding[1] much like the encoding levels in the CES guidelines. This would make it possible to develop a successive approximation to either end of the two different encoding strategies.

References

T. Erjavec, D. Tufis, and T. Váradi. Developing TEI-conformant lexical databases for CEE languages (CONCEDE project). In *In this volume*, 1999.

J. Juhász, I. Szőke, G. O. Nagy, and M. Kovalovszky, editors. *Magyar Értelmező Kéziszótár*. Akadémiai Kiadó, 1972.

C. M. Sperberg and L. Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford, 1994.

[1] We are grateful to Nancy Ide for this idea.

List of Participants

JÓZSEF ANDOR
Janus Pannonius University
7624 Pécs, Ifjúság útja 6., Hungary
andor@btk.jpte.hu

MARIANNE BERKA
Akadémiai Kiadó Rt.
1117 Budapest, Prielle Kornélia u. 4, Hungary
bem@akkrt.hu

FRANCESCA BERTAGNA
Istituto di Linguistica Computazionale, CNR, Pisa
Via della Faggiola 32, Pisa, Italy
f.bertagna@ilc.pi.cnr.it

ADRIAN CHIŢU
ICI, Bucharest
"Maresal Averescu", 8-10, 71316, Bucharest 1, Romania
chitu@rock.ici.ro

MARIE-HÉLÈNE CORRÉARD
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan, France
Marie-Helene.Correard@xrce.xerox.com

TOMAŽ ERJAVEC
Department of Intelligent Systems, Jozef Stefan Institute
Jamova 39 SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

CEDRICK FAIRON
LADL, Universite Paris
2 Place Jussieu 75005 Paris, France
fairon@ladl.jussieu.fr

ANA FERNANDEZ-PAMPILLÓN
Facultad de Filologia (Area de Linguistica) (Universidad Complutense de Madrid)
Departamente de Filologia Espanola I Faculdad de Filologia (Edificio B 7S planta)
Avda. de la Complutense s/n (Ciudad Universitaria) 28040 Madrid Espana, Spain
apampi@eucmax.sim.ucm.es

THIERRY FONTENELLE
European Commission Translation Service Development of Multilingual Tools
Jean Monnet Building JMO B4/77 L-2920, Luxembourg
Thierry.Fontenelle@sdt.cec.be

MAURICE GROSS
LADL, Universite Paris
2 place Jussieu, Cedex 05 Paris, F-75221, France
mgross@ladl.jussieu.fr

ULRICH HEID
Universitat Stuttgart, Institut für machinelle Sprachverarbeitung / Computerlinguistik
Azenbergstrasse 12, 70174 Stuttgart, Germany
uli@ims.uni-stuttgart.de

JACK HALPERN
Kanji Dictionary Publishing Society, UK
jack@kanji.org

ALISON HUETTNER
CLARITECH Corporation
5301 Fifth Avenue Pittsburgh, PA 15232 USA

PÉTER HUSZÁR
STEP Electronic Publishing Kft.
1113 Budapest, Bocskai út 77–79, Hungary

HITOSHI ISAHARA
Communications Research Laboratory, Ministry of Posts and Telecommunications
588-2, Iwaoka, Iwaoka-cho, Nishi, Kobe, Hyogo 651-2401, Japan
isahara@crl.go.jp

MAARTEN JANSSEN
ViL-OTS
Trans 10 3512 ED Utrecht, Holland
Maarten.Janssen@let.uu.nl

KYOKO KANZAKI
Communications Research Laboratory, Ministry of Posts and Telecommunications
588-2, Iwaoka, Iwaoka-cho, Nishi, Kobe, Hyogo 651-2401, Japan
kanzaki@crl.go.jp

FERENC KIEFER
Research Institute for Linguistics Hungarian Academy of Sciences
1014 Budapest Színház u.5–9, Hungary
kiefer@nytud.hu

GÁBOR KISS
Department of Corpus Linguistics,
Research Institute for Linguistics Hungarian Academy of Sciences
1014 Budapest Színház u. 5–9, Hungary
kissgabo@nytud.hu

JANA KLÍMOVA
Institute for the Czech Language,
Academy of Sciences of the Czech Republic
Letenska 4 118 51 Praha 1, Czech Republic
klimova@ujc.cas.cz

MATHIEU MANGEOT
Xerox Research Centre Europe
6 Chemin de Maupertuis, 38240 Meylan, France
Mathieu.Mangeot@xrce.xerox.com

ANTONIO MOLINA MARCO
Department of Information Systems and Computation
Camino de Vera s/n 46020, Valencia University of Technology, Spain
amolina@dsic.upv.es

OTTÓ TAMÁS MOLNÁR
Akadémiai Kiadó Rt.
1117 Budapest, Prielle Kornélia u. 4, Hungary
Molnar.Otto@akkrt.hu

KADRI MUISCHNEK
Tartu University
Ulikooli 18, EE-2400 Tartu, Estonia
kmuis@psych.ut.ee

CSABA ORAVECZ
Department of Corpus Linguistics,
Research Institute for Linguistics Hungarian Academy of Sciences
1014 Budapest Színház u. 5–9, Hungary
oravecz@nytud.hu

JÚLIA PAJZS
Department of Lexicography,
Research Institute for Linguistics Hungarian Academy of Sciences
1014 Budapest Színház u. 5–9, Hungary
pajzs@nytud.hu

TADEUSZ PIOTROWSKI
English Department, Opole University
Zielinskiego 47/11, 53-534 Wroclaw, Poland
tadpiotr@ii.uni.wroc.pl

BALÁZS POKORÁDI
Department of Lexicography,
Research Institute for Linguistics Hungarian Academy of Sciences
1014 Budapest Színház u. 5-9, Hungary
pokoradi@nytud.hu

GYÖNGYI POMÁZI
Akadémiai Kiadó Rt.
1117 Budapest Prielle Kornélia u. 4, Hungary
PG@AKKRT.HU

MARIONA SABATE
University of Lleida
Department of English and Linguistics Pl. Victor Siurana 1
25003 Lleida (Catalonia), Spain
msabate@dal.udl.es

JEAN SENELLART
LADL, Universite Paris
2 Place Jussieu 75005 Paris, France

MAX SILBERZTEIN
LADL, Universite Paris
2 Place Jussieu 75005 Paris, France
silberz@ladl.jussieu.fr

WOLFGANG TEUBERT
Institut für deutsche Sprache, Dept. for Lexical Studies
R 5, 6-13, 68161 Mannheim, Germany
Thompson@ids-mannheim.de

JAUME TIÓ
University of Lleida
Department of English and Linguistics Pl. Victor Siurana 1
25003 Lleida (Catalonia), Spain
jtio@dal.udl.es

DAN TUFIŞ
Romanian Academy (RACAI), Bucharest
13, "13 Septembrie", 74311, Bucharest 5, Romania
tufis@valhalla.racai.ro

TAMÁS VÁRADI
Department of Corpus Linguistics,
Institute for Linguistic Research Hungarian Academy of Sciences
1014 Budapest Színház u. 5–9, Hungary
varadi@nytud.hu

LÍDIA VARGA
Budapesti Műszaki Egyetem, Nyelvi Intézet / Universite Paris 7 LADL
Egry József u 1–3, Hungary
VARGAL@nyi.bme.hu

KADRI VIDER
University of Tartu
Department of General Linguistics
Tiigi 78 - 204 51003 Tartu, Estonia
kvider@psych.ut.ee

GEOFFREY WILLIAMS
Faculte des Sciences et des Techniques. University of Nantes
2 rue de la HOUSSI NIERE BP 92208-44322 Nantes CEDEX 3, France
william@ensinfo.univ-nantes.fr

ADNANE ZRIBI
University of Tunis III
I.S.G. de Tunis 41, rue de la liberte 2000 Le BARDO, Tunisia
adn@GNet.tn

CHIRAZ ZRIBI
University of Paris-sud, Centre of Orsay
Immeuble les orangers Appartement Nr.2 2080 Ariana, Tunisia
adn@GNet.tn